



北京·BEIJING

内 容 简 介

当产品经理遇上大数据时代，数据产品经理应运而生。新时代的新岗位自然也有新要求。数据思维、数据预处理、数据统计、数据挖掘、数据可视化等是产品经理的必备技能。懂产品、懂运营、懂市场、懂表达、懂管理则是数据分析师的技能外延。本书正是为有志于从事数据产品岗位的人士提供掌握上述技能的必修课。

让我们通过本书，在大数据的浪潮中乘科技与人文的扁舟，驶过数据产品经理的港湾，驶向数据科学家的彼岸。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

数据产品经理必修课：从零经验到令人惊艳 / 李鑫著.
—北京：电子工业出版社，2018.4
ISBN 978-7-121-33695-9

I. ①数… II. ①李… III. ①数据处理—产品设计 IV. ①TP274

中国版本图书馆 CIP 数据核字（2018）第 029472 号

策划编辑：孙奇俏

责任编辑：张春雨

印 刷：

装 订：

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱

邮编：100036

开 本：720×1000 1/16

印张：20.25

字数：350 千字

版 次：2018 年 4 月第 1 版

印 次：2018 年 4 月第 1 次印刷

定 价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

推荐序 1

我在中科大长期为本科生和研究生讲授人工智能相关课程，每次讲到人工智能的历史，就会提到人工智能之父——艾伦·图灵（Alan Turing），他也是英国著名的数学家和逻辑学家。二战时期，图灵曾协助军方破解了德国的著名密码系统“英格玛”（Enigma），从而扭转了战局，帮助盟军取得了二战的胜利。军事行动中破解密码的行为，其实就是在获取不对称的信息，从而占得先机，拔得头筹。

在和平年代，尽管没有烽火和狼烟，但商业环境中围绕信息的竞争却一点也不逊色于军事行动。掌握行业动态信息的基金公司可以把握股市的走向，从量化交易中获得利润；了解客户喜好的电子商务公司能够绘制用户画像，用精准推荐扩大流量；拥抱数据智能的政府机构可以打通数据孤岛，在海量数据指导下智慧治理。习近平总书记高屋建瓴总结性地指出“信息掌握的多寡成为国家软实力和竞争力的重要标志”“谁掌握了数据，谁就掌握了主动权”。

无论是量化交易程序，还是个性化精准推荐系统与智慧治理，都属于数据产品的范畴。数据除了在金融、电商、政府治理等领域有用武之地，其与实体经济也已深度融合，在各行各业都形成了增长点和新动能。然而利用好数据，做出适用于行业与特定领域的数据产品并非易事，这需要企业、机构员工都需具备一定的数据思维。

在我看来，正是由于海量的数据极大地扩展了人们的视野，传统的搜索才让位于个性化的精准推荐；正是由于数据快速的变化，人们才会目不暇接，甚至尚未从上一条数据所提供的信息中缓过神来，又要快速地投入到对下一条数据的处理。数据就这样从体量和速度上使得人们的认知有了盈余，从而我们可以在一个信息爆炸的时代站在历史的高度和宏观的尺度上，更为深入地理解、思考、判断。

我们为整个互联网贡献了语音、图像、文本、视频以及地理位置数据，作为回馈，也享受了互联网基于这些数据而为我们提供的全方位、多角度、便捷的服务。无论是从移动互联网时代走过来的数字移民，还是生长在万物互联时代的数字原住民，都具有一幅全息、多维的数字用户画像。参与得越深刻，画像越清晰。毫不夸张地说，现实世界的物理空间与网络世界的虚拟空间的界限越来越模糊，

数字化的自己已经逼近了真实的自己，数字化的事实也很有可能就是事实。多维数据的真实性越来越受到重视且不容置疑。

数据思维就像是现代化社会的一种方法论，它既是文化，也是工具；它既有阳春白雪的理论体系，也有下里巴人的实践指南。这一虚一实之间，顶天与立地之间，都润物细无声地透露着数据的价值。

本书作者曾在我的科大实验室中度过了 7 年的学生时光，学生时期他就善于深入浅出深奥知识，并分享给身边的人。这本书也延续了这个风格，原本需要高等数学基础的数据挖掘技术竟可以通过直白、简单的语言和比喻来说清楚，束之高阁的专业知识就这样走向大众，为更多的人服务。

书中详细介绍了大数据的来龙去脉，以及数据技能的方方面面。这既可以为企业中从事数据相关工作人员提供思维地图，也可以为其他机构中想要跨界了解数据世界的人提供一扇窗。另外，书中充满了关于数据历史、思维、技术周边的奇闻轶事，读起来想必不会枯燥，由此也可见作者广泛的阅读经历。

“纸上得来终觉浅，绝知此事要躬行”，要想发挥数据的价值，除了从思想上拥抱数据，更要掌握数据变换的规律，知晓数据挖掘的算法，运用数据产品的技巧，锻造数据人的自身修养，用数据指导实际工作。

希望《数据产品经理必修课：从零经验到令人惊艳》这本书能够带给你所需要的思维与技巧，也希望更多产业界的朋友能够做好合格的数据人，开发出优秀的数据产品，拥有惊艳的数据业绩！

陈恩红

中国科学技术大学教授

安徽省计算机学会理事长

安徽省大数据产业联盟理事长

推荐序 2

60 年前，苏联第一颗人造地球卫星被送入太空。作为冷战的响应，美国国防部组建了高级研究计划局（ARPA），开始将科技用于军事。11 年后，ARPA 组网连接了美国西部的若干个高校，成为了互联网的雏形。1983 年，美国国家自然科学基金会（NSF）投资建设了连接更多节点的广域网 NSFNet。1984 年 12 月，思科系统公司正式成立。

正如公司的标志性图标“旧金山大桥”所展示的那样，思科成立之初就致力于连接世界。从有线交换机到移动路由器，连接的自由增加了；从协作终端到在线视频会议，连接的内容丰富了；从互联网到物联网，连接的品类增多了。一切围绕连接，一切为了连接，一切走向连接，这似乎是技术发展亘古不变的主题。

在云计算时代，更多的物理资源被连接，而后数据孤岛的连接与打通又召唤来了蓬勃发展的大数据时代。数据的沉淀与喂养使得更多的智能算法被跨界连接在一起，成就了现在的人工智能时代。当下，每一个身处信息与通信行业的人都需要、也应该走出自己的小圈子，与周围更多的人和事物连接。

作为行业的典型，数据科学家是当今业界非常热门的岗位，但真正符合其职业要求的却少之又少。原因就在于数据科学连接了计算与管理，连接了技术与商业。跨学科、跨领域、跨门类的知识连接让很多人望而生畏。

“世上无难事，只要肯登攀”，数据科学家也许遥不可及，但数据产品经理却更容易上手。正如作者在书中所述：“数据产品经理既可以是拥有产品思维的数据分析师，也可以是掌握数据分析技能的产品经理。”只要能够做好数据技能与产品思维的连接，你就是一个合格的数据产品经理。这本书提供的就是这种连接。

我与作者相识 8 年，尽管当时“大众创业，万众创新”的号召还没有被提出，但还是学生的他就已经开始在宿舍“捣鼓”一个叫作“心愿 FM”的 App 创业项目。这款 App 提供类似校园社交的功能，其本质就是建立校园中异性之间的连接。时光荏苒，尽管这个项目已经不复存在，但作者却将“连接”的精神保留了下来，成为了一名跨界连接的专家。

阅读这本《数据产品经理必修课》，透过关于技术的字里行间，我看到了人文的斑驳倩影，不得不说，作者在写作上也将技术与人文进行了一次连接。

面向未来，我笃定地坚信，只有不断学习、自我更新的人才有可能胜出，成为企业需要的人才。学习既包括深度的探究，也包括广度的见识。无论是在大公司还是创业公司，是国企还是外企，了解一门新的知识总是有百益而无一害的，更何况是紧跟时代节奏的数据知识。

希望这本书能够真正帮助到有志于从事数据科学与数据产品工作的人，也希望它能成为读者和作者连接与沟通交流的纽带。

方剑斌

思科大中华区副总裁

前言

Preface

数据产品是什么

倘若我提出这样一个问题：数学中的 1、2、3 分别代表什么？你心中会有什么样的思考，又会有什么样的答案呢？早在 2500 年前的古希腊，毕达哥拉斯学派就已经给出了答案：点、直线与平面是对这三个数字的几何描述；源头、两性与稳定则是对它们深层次内涵的诠释。由此看来，数学有着神秘的意义。难怪古希腊数学家普洛克拉斯会说：“哪里有数学，哪里就有美。”

数据一词事实上是按照宾语前置的方式来构词的，所以我们可以理解为“据数”或“以数为据”，意思是把数据当成考究的凭证。正如数学是人类早期复杂贸易催生的结果，数据先天也带有商业的属性。从结绳记账到珠算发明，从证券股票到数字广告，我们甚至可以模仿先贤的口吻说道：“哪里有数据，哪里就有商业。”

历史的车轮已滚过千年，但数据的概念并未行将迟暮，垂垂老矣，反而老当益壮，焕发生机，这都要归功于“大数据”概念的产生。对于大数据这一概念，行业中有人将其归功于某家公司，有人将其产生与某位学者联系起来，但他们更多地是这个概念的精神作者，大数据真正的作者应该是接受并使用它的人，从这点来说，消费者才是其真正的衣食父母。

然而数据毕竟是一个虚构的概念，当我们谈论数据的时候，我们并没有办法在物理世界中找到一个实物来说明其客观存在。也许你会拿出刚打印出的报表，并反驳我说：“难道这不是数据吗？”可是你指的究竟是白色的纸还是黑色的油墨呢？由此可见，数据之名，需要借以载体之实才可以发挥价值，我们身边的产品就是这样一类载体。这么看来，打印出来的报表，包括邮寄上门的水电费单据，

都是数据产品。

数据产品的内涵应该不止这么肤浅，要不然岂不是“人人都懂数据产品”了？数据产品最为重要和关键的价值是驱使行动。水电费单据驱使我们缴费，推荐系统驱使我们阅读，财务指标驱使公司制定战略，就连菜市场用于标注商品价格的黑板也能驱使人们采购。如此看来，一个不能够驱使行动的数据产品其价值可能要大打折扣了。

让我们再次回顾一下数据产品这个概念，人的行动产生贸易，贸易产生数据，数据通过产品展现，数据产品驱动人的行动，人的行动又产生数据……周而复始，形成闭环。这才是一个完整的数据产品。

为什么要写此书

有关产品经理的著作有很多，仅在 2017 年，我拜读过的就不下 15 本。从琳达·哥乔斯到苏杰，从乔克·布苏蒂尔、卢克·米勒到陈峻锐、闫荣、后显慧、刘飞，我的产品思维便是从他们的荟萃中汲取的营养，因此你也许会从本书中读到他们的部分观点。关于大数据技术的书更是数不胜数，倘若将概念扩展到数据挖掘与机器学习领域，光是近几年出版的书籍便已不胜枚举。

如此说来，市场上似乎并不缺少有关数据产品的书，读者只需要兼读两者即可。那么为什么我又要写这本书呢？

著书并非我的本意，我的初心只是分享，这些内容最初被我写在我的个人头条号上，因为我信仰“分享是最好的学习方式”。在分享的过程中当然会受到质疑，但回答质疑乃至承认错误也是学习与进步的一部分。著书不过是分享的一种渠道罢了。

依我有限的阅读量和浅见来看，大多数技术类书籍充满了大量面向“圈内人”的专业术语，每一个术语都像是横梗在读者通往知识彼岸道路上的一座大山。并不是所有的读者都需要，或者愿意，甚至有能力“逢山开路”。“知识若庞杂到无法在民众中普及，则极易沦为经院哲学，甚至演化为民众对权威的盲目迷信”，威尔·杜兰特如是说。因此美国历史学家詹姆斯·哈维·罗宾逊号召“拆除壁垒，还知识于民众”，我仅仅是众多拥趸者之一罢了。

恰逢此时，我拜读了英国作家赫伯特·乔治·威尔斯（Herbert George Wells）的

《世界史纲》，这本装订成上下册的历史界的“红宝书”可以算是开了“概论”的先河，行文有趣，笔法生动，让我一个在历史方面十分愚钝的人也感起兴趣来。我也拜读了吴军老师与涂子沛老师几乎所有的著作，书中所介绍的科技背后的历史桥段让我反复琢磨。《吴晓波频道》和《罗辑思维》对我也颇有启迪。或许，将理性的数据用感性的故事进行呈现，是一种更易于让读者接纳的方式。我不禁这样想。

数据冷酷得像一个法官，故事却像富有温情的妇人，理性与感性的矛盾不言而喻。不仅如此，科技与人文的较量，非虚构与文学的角力，也都正在进行。尽管美国作家菲茨杰拉德告诉过我们：“测验一个人的智力是否属于上乘，只看脑子里能否同时容纳两种相反的思想，而无碍于其处世行事。”但知易行难。

为了获得这样的智慧，我们需要找到途径和方法。好在查理·芒格、赫伯特·西蒙（司马贺）以及小泉英明各自都著书立说，为我们提供了工具与方向。这是一种被称为多学科交叉的思考方式，我也希望用这种方式来写作与分享，让这本书在丰富内容之余兼具有趣的灵魂。

本书内容

全书分四大部分，共计 15 章，每部分及每章的具体内容如下。

第一部分 产品经理的前世今生

第 1 章 产品经理的前世

产品经理一词究竟是何意义？该岗位从何而来？广义与狭义上有何区别？当前互联网行业的产品经理究竟做些什么？本章将为你一一揭晓。

第 2 章 产品经理的今生

提出管理动机、广义竞品分析与交互设计这样的对产品经理的更高要求。另外，本章也将梳理产品经理在产品、团队、公司层面必须迈过的雷池。

第 3 章 产品经理的入行

“宽进严出”是产品经理岗位的特色，低门槛使得初入此行甚是轻松，高要求则使得出类拔萃愈加困难。本章将介绍从事产品经理的人员该具备什么样的视野，又该在工作中如何学习。

第二部分 古往今来的数据思维

第4章 历史中的数据思维

数据埋点、数据可视化、数据产品落地、数据驱动决策、利用数据降本增效、统计分析以及打通数据孤岛这些老生常谈的话题似乎可以在过往的岁月中找到关联。本章向历史溯源，给出老概念的新故事，诉说新时代的旧往事。

第5章 行业拥抱数据思维

从蓝色星球到 960 万平方公里的泱泱大国，从与政府密切联系的科教文卫体到与民生息息相关的衣食住行，本章将介绍大数据渗透到的每一处角落。

第6章 当产品经理遇见数据思维

当产品经理遇见数据思维的时候，不仅有“眼前的苟且”——数据产品经理，还有“诗和远方”——数据科学家。从现实到理想的距离，本章将为你搭梯。

第三部分 数据产品经理的技能进阶

第7章 面向产品经理的数据预处理

“磨刀不误砍柴工”，做好数据预处理可以为数据分析与挖掘过程节省许多时间。本章从数据清洗、数据集成、数据变换以及数据规约四个角度全面阐述数据预处理的相关知识。

第8章 面向产品经理的统计分析

本章从非时序数据与时序数据、分类数据与连续数据的角度，介绍数据统计与数据分析的概念及技巧。

第9章 面向产品经理的数据挖掘

本章旨在讲清楚数据挖掘的方方面面，内容包括回归、分类、聚类、关联分析和时间序列分析等数据挖掘算法，以及在此基础上的集成学习、文本挖掘、社交挖掘、排序算法、推荐系统以及用户画像。最后以这些算法中蕴含的哲学内涵作为结束。

第10章 面向产品经理的数据可视化

数据可视化是“技术与美”的最好结合。本章分别介绍“高大上”与“接地

气”的两类数据可视化，并从数据展现和逻辑修饰两个层面介绍识别数据“说谎”的技巧，最后给出数据可视化的终极形式——数据报告的制作方法。

第 11 章 向数据科学家再迈一步

本章介绍与数据产品岗位相关的另外三个岗位，分别是运营、研发与市场。在通往数据科学家的铸鼎之路上，这三足必不可少。

第四部分 数据产品经理的自我修养

第 12 章 学习力：借方法论加速

本章系统介绍工作中遇到的各种方法论，并概括总结诸多方法论的“模板”，最后提出笔者自己的学习方法论，帮助读者建立工作中的“理论自信”。

第 13 章 表达力：用逻辑学帮衬

本章从“为国考正名”谈到“名著中的名言警句”，详细介绍产品经理日常工作与案头写作的心得，以及汇报与分享中极具感染力的“故事思维”。

第 14 章 领导力：以经济学诠释

本章在经济学原理中找到管理学中领导力的跨学科基础——无论是团队制定目标时的“举旗定向”，还是实际工作时的“谋篇布局”；无论是团队内部配合的“取长补短”，还是团队之间协作的“互利共赢”。

第 15 章 软实力：靠心理学打造

本章通过心理学中的若干现象与小实验，介绍帮助数据产品经理灵活驾驭本职工作的若干心得与技巧。

联系作者

无论是随意翻阅至此，还是计划花一番功夫细细研读，从你捧起此书的那一刻起，我相信你我之间便产生了神奇的联系与耦合。书，就是我们之间的桥梁与纽带。我是本书的唯一作者，这意味着所有书中可能存在的疏漏与谬误也尽归于我，对此我表示歉意并将竭力修正。而我却又不是本书的唯一贡献者，出版社的编辑、公司的同事、身边的朋友以及读者朋友们的每一条建议都帮助过我进步，使这本书更加精细。

欢迎通过我的微信 `ustclixin` 与我取得联系，我会尊重大家的每一条意见和建议，并愿意与大家一起探讨有关数据产品的任何话题。

无以名状的感谢

尽管我是此书是唯一作者，但成书并非我一人所能为。儿时囫圇吞枣习得的典籍，现在不求甚解读下的篇章，是家庭给予我最宝贵的精神财富。

数据思维的启迪要归功于中科大的陈恩红教授，同时，实验室为我提供了徜徉在数据海洋的一叶扁舟。

科大讯飞公司中的一些同事参与了本书早期内部培训版本的审校，头条号与 PMCAFF 社区中的网友也向我提出了极具思考性的问题，在此不详细列举他们的姓名，一并谢过。

尽管成书早期也接触了其他出版社和编辑，但最终还是选择了博文视点，选择了给我提出建议最多，需要我动笔修改最多的孙奇俏老师，因为我相信她的认真与执着会让我更加专业。

关于内容，我如履薄冰，无论是鼓励，还是建议，抑或是批评，我都一并感谢并将铭记，且承诺会认真修正。

最后，再次深深感谢家人！

李鑫

2018 年 2 月于合肥

读者服务

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **提交勘误：**您对书中内容的修改意见可在 [提交勘误](#) 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动：**在页面下方 [读者评论](#) 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：<http://www.broadview.com.cn/33695>



目 录

Contents

第一部分 产品经理的前世今生

第 1 章 产品经理的前世

1.1	产品经理究竟是什么	4
1.1.1	咬文嚼字说产品经理	4
1.1.2	产品经理的历史溯源	5
1.2	泛产品经理与产品经理	6
1.2.1	产品经理的专业取向	7
1.2.2	产品经理的泛化	8
1.3	互联网产品经理的规定动作	12
1.3.1	需求调研	12
1.3.2	竞品分析	14
1.3.3	原型设计	16

第 2 章 产品经理的今生

2.1	卖家秀：自我提升的几项技能	20
2.1.1	从需求文档到动机文档	20

2.1.2	从竞品分析到广义竞品分析	22
2.1.3	从原型设计到交互设计	24
2.2	买家秀：弄垮团队的若干“要领”	28
2.2.1	越过产品雷池	28
2.2.2	踏入团队雷池	29
2.2.3	迈向公司雷池	30

第3章 产品经理的入行

3.1	入行做产品的几种可能	34
3.1.1	源自技术岗	34
3.1.2	源自业务岗	35
3.1.3	源自应届生	36
3.2	上岗后的第一件事	37
3.2.1	产品全图	38
3.2.2	行业全图	39
3.2.3	产业全图	40
3.3	工作中如何学习	41

第二部分 古往今来的数据思维

第4章 历史中的数据思维

4.1	人口普查：最早的数据埋点策略	46
4.1.1	埋点的技术视角	46
4.1.2	埋点的时机与策略	48
4.2	命令与征服：可视化最早的用意	49
4.2.1	可视化大家说	50

4.2.2	可视化与历史.....	51
4.3	科技革命：助力数据产品落地.....	54
4.3.1	手工统计.....	55
4.3.2	机械统计.....	55
4.3.3	电子统计.....	57
4.4	数据驱动决策的历史溯源.....	57
4.4.1	美国建立时用数据分权.....	58
4.4.2	南北战争时用数据进军.....	59
4.4.3	经济发展时用数据裁判.....	60
4.5	管理咨询：使用数据降本增效.....	61
4.5.1	咨询指引数据产品方向.....	62
4.5.2	管理启迪思维模式更新.....	63
4.6	聊聊统计学.....	64
4.6.1	政治算术.....	64
4.6.2	频率学派.....	65
4.6.3	概率学派.....	66
4.7	LEHD：美国的第一个大数据项目.....	67
4.7.1	信息逐步开放.....	67
4.7.2	大数据项目开展.....	68
4.8	历史给我们数据思维的启示.....	69
4.8.1	用数据说话.....	69
4.8.2	向贤者取经.....	69
4.8.3	渐进性创新.....	70
4.8.4	需求创造供给.....	70

第5章 行业拥抱数据思维

- 5.1 大数据从何而来..... 72
 - 5.1.1 大数据历史 73
 - 5.1.2 自身发展 75
- 5.2 大数据的全球格局与中国面貌 76
 - 5.2.1 全球格局 76
 - 5.2.2 中国面貌 77
 - 5.2.3 行业概览 78
- 5.3 大数据+ “治理与交通” 81
 - 5.3.1 治理 81
 - 5.3.2 交通 83
- 5.4 大数据+ “零售与金融” 84
 - 5.4.1 零售 84
 - 5.4.2 金融 88
- 5.5 大数据+ “体育与教育” 89
 - 5.5.1 体育 89
 - 5.5.2 教育 91
- 5.6 大数据+ “医疗与旅游” 93
 - 5.6.1 医疗 93
 - 5.6.2 旅游 94
- 5.7 大数据+ “农业与制造” 96
 - 5.7.1 农业 96
 - 5.7.2 制造 97
- 5.8 大数据行业成熟了吗 97
 - 5.8.1 行业成熟度 98
 - 5.8.2 大数据理念 99

5.8.3 大数据趋势	100
5.9 大数据在产业中的位置	103
5.9.1 行业组成	104
5.9.2 产业构成	106

第6章 当产品经理遇见数据思维

6.1 下一站：数据科学家	110
6.1.1 数据科学的历史由来	110
6.1.2 数据科学与商业智能	111
6.1.3 数据科学的职业分类	112
6.1.4 数据分析的技能进阶	114
6.2 数据产品经理的职业新要求	115

第三部分 数据产品经理的技能进阶

第7章 面向产品经理的数据预处理

7.1 数据分析的标准姿势	128
7.2 淘洗数据沙砾（数据清洗）	130
7.2.1 缺失值	130
7.2.2 异常值	132
7.2.3 归一化	133
7.3 聚细沙成佛塔（数据集成）	135
7.3.1 实体识别	135
7.3.2 冗余性识别	136
7.4 换个姿势再来一次（数据变换）	137
7.4.1 离散化	137

7.4.2 属性构造	139
7.5 少即是美（数据规约）	139
7.5.1 特征规约	140
7.5.2 样本规约	141

第8章 面向产品经理的统计分析

8.1 说有信息量的话（非时序数据的统计量）	144
8.1.1 集中趋势	145
8.1.2 离散趋势	146
8.1.3 数据分布	148
8.2 股票指数是什么（时序数据的统计量）	148
8.2.1 “三比”	149
8.2.2 股票指数	150
8.3 男女真的有别吗（分类数据的统计量）	152
8.3.1 卡方是什么	152
8.3.2 卡方怎么算	153
8.4 相关性不是因果性（连续数据的统计量）	156
8.4.1 Pearson	156
8.4.2 Spearman	157
8.4.3 Kendall	158
8.5 数据不能承受之“熵”	159
8.5.1 物理中的“熵”	159
8.5.2 信息中的“熵”	160

第9章 面向产品经理的数据挖掘

9.1 学数据挖掘，只需要高中数学	164
9.1.1 重温“加减乘除”	164

9.1.2	重温“比值”	165
9.1.3	重温“函数”	165
9.1.4	重温“符号”	165
9.2	线性回归：人为什么没有严重两极分化	166
9.2.1	优生学趣闻	166
9.2.2	空间中的直线	167
9.3	逻辑回归：种群增长的 S 型曲线	169
9.3.1	种群的增长曲线	169
9.3.2	S 型曲线的秘密	171
9.4	朴素贝叶斯：面相占卜工作原理	172
9.4.1	外貌协会与街头看相	173
9.4.2	无处不在的贝叶斯	174
9.5	决策树：爱情选择背后的心理学意义	176
9.5.1	爱情选择条件多	177
9.5.2	不纠结的小技巧	178
9.6	K-means：寻找物理学上的质心	181
9.6.1	向中心看齐	181
9.6.2	站错队的后果	183
9.7	层次聚类：分而治之与抱团取暖	184
9.7.1	分而治之	185
9.7.2	抱团取暖	185
9.8	DBScan：帝国崛起的定居、建国与扩张	186
9.8.1	密度打败划分	187
9.8.2	相似的帝国发展路径	188
9.9	关联规则挖掘：“啤酒和尿布”是个谎言	188
9.9.1	讹传已久的商业故事	189

9.9.2 关联规则的三重门	190
9.10 时间序列分析：聊聊《周易》	192
9.10.1 时间序列分析的玄妙	192
9.10.2 时间序列分析的正经	194
9.11 集成学习：三个臭皮匠赛过诸葛亮	195
9.11.1 多拜师与拜大师	196
9.11.2 向大家与失败学习	197
9.12 文本挖掘：让机器读懂你	199
9.13 社交网络：隐私无处遁形	202
9.14 排序：简约而不简单的事	205
9.14.1 排序的规则方法	205
9.14.2 排序的操作机理	207
9.15 推荐系统：“今日头条”背后的秘密	208
9.16 用户画像：隐私是个“伪命题”	213
9.17 算法思想中的哲学内涵	216

第 10 章 面向产品经理的数据可视化

10.1 别人家的可视化：阳春白雪	222
10.2 工作中的可视化：下里巴人	227
10.3 用可视化“说谎”	230
10.3.1 数据的误导	230
10.3.2 逻辑的谬误	234
10.4 准备一份数据报告	238

第 11 章 向数据科学家再迈一步

11.1 能文：陪运营跟踪产品看效果	244
11.1.1 传统运营的基本功	245

11.1.2	数字化运营“三”话你知.....	248
11.2	能武：追研发把控进度出成果.....	251
11.2.1	数据采集.....	251
11.2.2	数据存储.....	254
11.2.3	数据计算.....	256
11.2.4	数据分析.....	258
11.3	能聊：跟随销售面向市场找思路.....	258

第四部分 数据产品经理的自我修养

第 12 章 学习力：借方法论加速

12.1	方法论知多少.....	266
12.1.1	概念阐述.....	266
12.1.2	分类总结.....	267
12.2	学习过程的“满灌”与“脱敏”.....	269
12.2.1	理解提炼.....	269
12.2.2	我的方法论.....	271

第 13 章 表达力：用逻辑学帮衬

13.1	写得一手好文案.....	274
13.1.1	为公务员考试正名.....	274
13.1.2	写作实战简明教程.....	275
13.2	讲故事给同事听.....	278

第 14 章 领导力：以经济学诠释

14.1	事情背后的选择.....	285
------	--------------	-----

14.1.1	选择价值链上游：剪刀差效应.....	285
14.1.2	学会审时度势：美林时钟.....	286
14.1.3	谨慎选择别人的经验：推绳子效应.....	286
14.1.4	平衡是一个难题：萨伊定律与凯恩斯法则.....	287
14.2	人员之间的协同.....	288
14.2.1	你闪开，让我来：绝对优势与相对优势.....	288
14.2.2	无条件开放：零和博弈与合作共赢.....	289
14.2.3	教会团队成员什么是沉没成本.....	290

第 15 章 软实力：靠心理学打造

15.1	向内求：耐心、谦逊、热心.....	294
15.1.1	让自己“延迟满足”.....	294
15.1.2	对表扬免疫.....	295
15.1.3	不怕丢脸地分享.....	297
15.2	对外看：大局、妥协、有趣.....	297
15.2.1	看问题需要“上帝视角”.....	298
15.2.2	率真对内，圆滑对外.....	298
15.2.3	一切从简，有趣有梦.....	299

part one

01

第一部分

产品经理的前世今生

第1章

产品经理的前世

- 1.1 产品经理究竟是什么 4
 - 1.1.1 咬文嚼字说产品经理 4
 - 1.1.2 产品经理的历史溯源 5
- 1.2 泛产品经理与产品经理 6
 - 1.2.1 产品经理的专业取向 7
 - 1.2.2 产品经理的泛化 8
- 1.3 互联网产品经理的规定动作 12
 - 1.3.1 需求调研 12
 - 1.3.2 竞品分析 14
 - 1.3.3 原型设计 16

1.1 产品经理究竟是什么

都说物物皆产品，人人都是产品经理，那么这个话题显然是一个大众话题，至少在互联网领域中如是也。正如“一千个人眼中有一千个哈姆雷特”，每个人关于产品经理的理解也应该有所不同。我从微观与宏观两个角度，分别做了“抠字眼”与“找历史”两件事，希望说明我心中的产品经理究竟是什么。

1.1.1 咬文嚼字说产品经理

谈到产品经理，必然要谈到产品，可究竟什么是产品？不妨让我们先来咬文嚼字一番。

《说文解字》（以下简称《说文》）是一部系统地分析汉字字形、考究字源的字书，该书对产品二字的解读颇具启示。《说文》中对于“产”字的描述是“产，生也”，而对于“品”则描述成“品，众庶也”。该怎么理解这样的讲法呢？说到“产”，我们最容易联想到的词汇是“生产”“出产”等，表达的意思大致是生长与制造，分别对应了农业与工业从无到有、从零到一的场景。而对于“品”，其本身是由三个口组成的，口代表人，三代表多，因而品的原意是众多的人，这也是《说文》一书中给出其“众庶”含义的依据。然而在汉字的使用过程中，品字逐渐具有了表示等级、性质、评判等的引申义，如品色、品质、品味等，这是因为其过程中的断定与判断离不开人的参与。

综合上面对于“产品”二字的注释，我们大致可以把产品的生命历程分为两个过程，一个是制造生产的过程，一个是人参与消费与评价的过程，这和产品经理所参与的业务流程也基本吻合，粗浅说来就是面向内部的产品规划、设计与研发，以及面向外部的营销、销售、运营与售后服务。

关于经理一词，早在战国末年就已经出现。赵国人荀况和其弟子所著《荀子·正名》中提到：“心也者，道之工宰也；道也者，治之经理也”，意思是，治理国家的常法规矩就是道的反映，而主宰道的是人的内心，即心中的想法、信念、情绪与心态。唐朝人杨倞在《荀子注》中对于经理二字做了诠释，注解中说道：“经，

常也；理，条贯也”，意思就是，经理二字表明了经常性的条例。

联系现如今的经理工作，特别是职业经理人处理的日常事务，不能不说是一种经常性的条例，这种经常性主要体现在两处：一是处理的频率，处理事务往往就是经理平日最经常性发生的动作；二是处理的规范性，经理，特别是职业经理人，基本上都是基于一套标准化的流程来办事，这就是条例的作用。从这个角度来说，产品经理就是围绕产品经常性地执行一些标准化动作（如市场分析、需求调研等）的一类人。

产品经理的英文名称是 Product Manager (PM)，许多网站也以此作为产品经理社区的识别符号（如人人都是产品经理社区，woshipm.com）。可是 Manager 这个词译为经理，却似乎是一种既准确又在特定语境下容易产生误解的折中办法与妥协。一方面，说它准确，是因为在美林韦氏词典（Merriam-Webster Dictionary）中，Manager 一词的释义中充满了 conduct、direct 等词汇，意为实施、组织、指导、控制的人，对应到产品经理的角色上，产品经理的确对产品具有指导和组织的职责，并成为产品的第一负责人，控制产品生命周期的各个阶段。另一方面，经理一词在职场上有特定的含义，它的意思是管理者，即有一定职权的人，如公司的总经理。然而产品经理大多数情况下并不是一个经理，在矩阵式管理的企业中，职能线与业务线交错纵横，产品经理往往没有直接下属，也没有权限要求其他人向自己汇报，更多的是协调与沟通。从这个角度来说，产品经理更多靠的是无授权领导或者个人魅力，而不是职位的势能来推动工作。这也对产品经理提出了更高的要求。

1.1.2 产品经理的历史溯源

产品经理是互联网行业中的流行词汇，其岗位需求量也与日俱增，然而这并不是一个互联网行业的专属词汇。工业产品伴随着 18 世纪英国工业革命的兴起而出现，如果继续向前追溯，农业产品理所应当地成为产品的最早期形态。而在农业与工业社会中，也存在着某种“产品经理”。

农业文明的社会未必会称农业产品的负责人为产品经理，毕竟经理一词伴随着现代公司制度的产生而普及。在农业社会，人们经营的往往是农产品或者农业

加工品生意，那些有着一技之长，独立或协助打理生产及采购流程，执行营销或接待任务的人则成为了农业社会的经理，在古时候扮演这样角色的人或是手艺人，或是能带上几个徒弟，被人称为“师傅”的人，有些类似于今天互联网公司中的 tech-leader（技术领导者）。还有一些人在工农业交杂的时期受雇于商家与店铺，进行经营活动，这样的群体在中国古代有一个响亮的名称——“掌柜”，这有些类似于职业经理人，而其雇主则更像是现代公司制度中的董事长。

18 世纪下半叶，工业革命出现，劳动力得以解放，工场手工业转变为机器大工业，进而促进了工业与农业的分离。从那时起，工业历经蒸汽机械、电气化、信息技术这三个阶段，才到今天的智能化阶段。德国政府提出的“工业 4.0”计划，以及我国提出的“中国制造 2025”均为这个阶段的工业发展提出了清晰的路线图。技术越是发展，经济越是繁荣，这已经是不争的事实，随之带来的还有新的社会分工现象。产品经理就是在工业革命的过程中伴随社会分工而出现的。

分工的过程可以概括为“大则分，分须立”。一方面，只有当一些工作内容或技能集中于一个人身上达到不可兼顾进而影响工作质量的地步时，才需要进行拆分，即满足了“大到一定程度才拆分”的原则。另一方面，拆分出来的部分技能的经济收益足够养活只拥有这批技能的人，也就是说这样的技能在市场中有广泛的需求，即保证“拆分出来的拥有这项技能的人能够立足”。

公司也是工业时代的产物，其生命周期中从成长到壮大的过程，也是分工的最佳印证。粗浅地说，公司中聚集着一批不同领域的专业人士或手艺人，靠着彼此间的协作，在市场上从事经营活动，而经理则是确保协作顺利进行且对协作结果负责的人。当公司发展壮大并开始进行多目标经营时，每一个目标项目都应当配备一个经理，如同细胞分裂一样，当细胞的遗传物质进行复制并开始膨胀时，便需要进行拆分，此即事业部的雏形。而当经理所管辖的工作已经应接不暇时，也就诞生了事业部中的各种经理，如产品经理、营销经理、人事经理等。

1.2 泛产品经理与产品经理

各行各业都有产品经理，他们虽岗位要求各不相同，但同时共享着一些技能。当目光聚焦到一个垂直领域的时候，我们可以清晰地了解该领域产品经理的技能

细节；而当我们把各行各业的产品经理放在一起进行“眯眼测试”的时候，似乎又可以把握到一些大体相同的轮廓。前者被称为专业的产品经理，而后者被称为泛化的产品经理。

1.2.1 产品经理的专业取向

泛产品经理的概念，我最初是在苏杰老师的《人人都是产品经理 2.0》一书中看到的。这里所说的泛，以我的理解至少有两层含义，一是能力的泛化，二是领域的泛化。能力的泛化是指，要具有更为全面的互联网产品经理的能力，除去需求、原型与运营等，还需要了解商业模式、财务常识等。对于领域而言，产品经理的岗位不再，不应该，也从来没有局限于互联网或软件领域。每一个领域都需要产品经理，至少从事产品、品牌工作的人都应该具有产品经理的思维。我想这也是现在很多行业都喊出“人人都是产品经理”或者“产品思维”的原因所在。

有泛化，就一定有与之对应的窄化、聚焦与细分。相对于生产制造业中的产品经理，互联网产品经理就是一种聚焦与细分。美国威斯康辛麦迪逊大学商学院职业能力发展中心（Center for Professional and Executive Development, CfPED）的琳达·哥乔斯（Linda Gorchels）是一位畅销书作者，她的作品 *The Product Manager's Handbook* 与 *The Product Manager's Field Guide* 被翻译成中文，也就是大家熟知的《产品经理的第一本书》与《产品经理的第二本书》。这两本书首次出版距今已经近 20 年，书中的案例更多选自工业与消费品企业，其中一个典型案例就是，创建近 200 年的宝洁公司（P&G）为了减少广告浪费与顾客遗漏，于 20 世纪 30 年代初为两个独立的品牌香皂指定不同负责人的制度创新，形成自有品牌相互叫阵与竞争的态势，进而提振业绩，引得消费品企业竞相效仿。这也算是给产品经理开了一个好头。

很显然，琳达的书并不是一本专门写给当下互联网产品经理看的书，更不会教授互联网产品经理工作的实务与技巧。正如琳达在书中开头所说：“产品经理适用于所有企业，包括消费性包装企业、服务业、工业性产品企业以及非营利组织（如医疗机构等）”。书中的对象也是现在被我们称为传统行业的品牌经理们。这些品牌经理从诞生之初就要关注市场、生产、营销等，在当下看起来，这更像是

泛产品经理的职责，而他们却仍然被称为产品经理。

那么究竟是先有了产品经理还是先有了泛产品经理呢？

在《产品经理的第一本书》的最后一章中，琳达给出了产品管理者的三个发展方向，其中一个就是“更为专业的取向”。由于企业分派给产品经理的职责越来越多，对于产品经理一个人来说，工作负荷过重，按照我们之前提出的分工的两个原则，产品经理便进行了更为专业的分工，一种产品经理注重于消费者的议题，而另一种则更为关注贸易商的议题。特别地，在如互联网公司等高科技企业中，产品经理则专心处理产品在工程和技术方面的问题，而把大部分营销工作交给另外的职能单位来负责。

读到此处，答案已经不言而喻，苏杰老师所说的泛产品经理实际上应该是琳达所指的产品经理，而我们日常接触的互联网产品经理及其职责，则是产品经理在互联网专业领域的细化与分工。所以在我们所从事的互联网或 IT 软件领域，谈到产品经理，实际上更多是指互联网产品经理。

本书所介绍的数据产品经理也是一种分工与细化，可以是具备一定数据思维和数据分析能力的产品经理，也可以是具备产品思维的数据分析人员。

时过境迁，产品经理这一名词的高地已经被互联网领域人士拿下，而为了与之区分，传统行业则使用了品牌经理、行业经理等名词来称呼其产品负责人。随着产品经理能力的细化，产品经理这个岗位会向更为专业化的方向演进。

1.2.2 产品经理的泛化

狭义上来说，产品是商业时代的产物，产品经理自然存在于商业世界。从更为广义的角度来看，产品与产品经理的概念可以被泛化，产品经理实务型技能（如财务与经济常识、市场与竞争分析能力、营销与策划能力等）皆可以作为指导个人和组织在面对选择时做出正确判断的准则及行动指南。

在我看来，产品经理工作中的一些能力可以从职业中剥离出来，作为一般性的原则，我将这些原则称为 CAROL，意思是圣诞时欢快的颂歌，暗示着当我们掌握了这些一般性的原则之后就可以游刃有余地驾驭日常事务，并像圣诞节时那样欢快地 cherish（珍视）工作与生活。CAROL 中的每个字母都代表一种普适能力。

闭环（Closed-loop）

闭环的思想在商业中已经被广泛采纳，无论是 SocialCRM 的闭环营销理论，还是 Six Sigma(六西格玛) 的管理模型，无论是精益分析中的 AARRR 海盗模型，还是标准数据挖掘强调的 CRISP-DM 流程，这些都是将闭环的思维方式应用到各个领域的实例。模拟电路中也有闭环的概念，这里的闭环指的是反馈回路或反馈控制系统，即把系统的实际输出和我们的期待输出进行比较，通过两者的偏差进行调节，使得实际输出接近于期待输出。

闭环不仅仅在系统理论与控制理论上被论证可行，在实际生活中也有很多印证。政府的信访机构与纪委部门对待群众来信时往往会遵循一个原则——“件件有着落，事事有回音”，这其实也是一种闭环。有人递交一份文件或者问询函，或是通过电话进行查询与咨询，是否给予反馈与落实，明则表现出一个人的处事态度，实则考究一个人或组织的靠谱程度与执行力。

了解闭环并使用它，做到“件件有着落，事事有回音”，将极大程度地提高他人对自己的信任以及系统的稳定程度。

抽象（Abstraction）

在小学的语文课本中有这样的习题：将一些词语放在一起，然后写出一些和这些词语相近的词语。如题面是“鸡、鸭”，那么出于它们都是家禽的考虑，可以填写“鹅”，或者出于它们都有羽毛的考虑，可以填写“燕子”。美国研究生入学考试（GRE）中也有这样一类题型，除此之外，国家与地方公务员考试中也有类似的试题，其本质上考察的都是逻辑抽象的能力。

抽象能力是一种总结和归纳的能力，也是一种透过现象看本质的能力。当我们把鸡、鸭抽象为家禽的时候，我们就获得了一类事物的某种属性。当我们把这个过程反过来，即通过抽象概念的家禽再去找一些实例（如鹅）的时候，我们就在进行演绎。简单来说，抽象与归纳是由特殊到一般，而演绎是由一般到特殊。

产品经理的日常工作中需要针对多个用户访谈或观察实例进行用户需求的理解与筛选，简单剖析是在产品功能层面，而深层次的剖析则是在心理诉求层面，即透过现象看本质。这样的抽象能力对我们而言至少有两个好处，一是在广度上帮助我们找到事物之间的联系，联系越多，思考越多，越容易产生创意、点子；

二是在深度上帮助我们挖掘背后的原因，越是不停地追问与抽象，就越有可能知道真相，行动也越踏实。

试错（tRail & eRRoR）

2016年11月28日，人民日报发表评论员文章《敢于试错是一种改革智慧》，文章开头引用奥地利批判理性主义哲学家波普尔的看法，即把科学看成“从错误中学习”，然后悉数了从小岗村大包干的“试”到邓小平同志关于股市与证券的“试”，从加入世贸组织对外开放的“试”到国企改制与供给侧改革的“试”，从而建设性地提出中央全面深化改革领导小组关于试错的态度——“既鼓励创新、表扬先进，也允许试错、宽容失败”，同时从另一侧面批判了由于不作为而导致的试错空间的关闭。

苏联科学家根里奇·阿奇舒勒（Genrikh Altshuller）也提出了一种被称为 TRIZ（发明问题解决理论，另一英文译名为 TSIP）的创新理论，其主要观点是，解决复杂系统中出现的难题、冲突与矛盾是推动整个系统进化的不竭动力，这算是对试错理论的发展。也就是说，不可能在行动之前就规划好一切，只有在行动后遇到困难和错误，不断解决与动态调整，才能使整件事与整个系统向前演进。

产品经理在日常工作中使用的精益思想、最小可用产品（MVP）理论，以及小步快跑、快速迭代等主张便是试错法最好的证明。我们不可能期望一整套的解决方案都被完美执行，我们需要的仅仅是明确目标，迈进一步，动态调整。

开放（Open-mind）

把目光转移到生物界，会发现一个有趣的现象，生物的体型千差万别，仿佛是一个漏斗的两端。在漏斗的一侧，如菌落、蝼蚁、珊瑚，它们是小个子的代表；而在另一侧则是大象、鲸鱼等巨兽，它们是大个头的示例。大个头往往带给我们沉重、垂暮、笨拙之感，而小个子则是灵活、轻便、旺盛的代名词。它们生存的方式也不尽相同，大个头们倾向于自给自足，而小个子们则更多依赖与同类协作完成任务而存活。当巨变来临时，大个头们往往面临“创新者的窘境”，在劫难逃，恐龙就是最好的例证，而小个子们则由于各种优势得以苟活。在我看来，小个子们具备了开放的特征。

开放有两种形式，一种是补给式，另一种是索取式。补给式是将自己的能力

开放给外界，从而获得与外界的联系，而索取式则是从外界获取营养与信息，而与外界建立关联。这与互联网中节点的出度与入度如出一辙，它们共同对这个节点在网络中的重要性产生影响，谷歌公司的网页排序算法 PageRank 便利用了这个特性。

在生活中，我们向什么人学习，以及影响什么人，直接决定了我们作为“产品”的品质优劣；而在产品生产过程中，我们运用什么技术与方法，以及生产的产品能够为环境提供多少价值，则决定了产品生命周期的长短。

学习（Learning）

这是一个熟悉的陌生词汇。说它熟悉，是因为成长在义务教育制度下的我们对它不陌生，随着社会和经济的发展，人均学习和受教育的时间越来越长。说它陌生，是因为我们对于学习的概念、内容、方法、层次却鲜有认知。

从概念上来说，学习的地点不再局限于校园、课堂以及培训班等封闭场所，突破了空间的学习，在时间上也不断突破人们可接受的底线，终身学习的概念已经不是空谈了。

从内容上来说，学习大体上可以分为洞察、体系、案例与娱乐四种。它们既有联系也有不同，联系是，这些内容共享了几乎所有的学习呈现形式，唯一的区别是学习者如何界定和理解它们。洞察是那种“一语点醒梦中人”的学习内容，是精神；体系是最便捷地进入新领域的完备的学习内容，是骨骼；案例是用来佐证洞察和体系的学习内容，是肌肉；娱乐是让我们调节节奏的学习内容，是装饰。

从方法上来说，模仿本身就是一种学习形式。著名经济学家林毅夫先生曾说过，独特并不是独一无二，而更多的是一种综合，一种借鉴。他也曾列举清代乾隆年间的学者纪晓岚编书不著书的示例来说明，广度上的模仿与综合本身就是一种学习，一种创造。

从层次来说，学习分为知道、行动、授予与变通四个层次。即包含了从知行合一的层次到传道予人，再到灵活变通的转变。

产品的自我学习就是产品的迭代，而产品经理的自我学习则是关于市场、用户、产品的认知升级。每当自己“脱一层皮”，就意味着自己在某个技能上又精进了一点，手艺也更进一步。

无论是社会环境、正式或非正式的组织，乃至个人，都是一个产品。组织的掌舵人以及自我经营的个体都需要具备产品经理的思维来看待产品这个系统。组织与组织的竞争以及个人的精进可以使用市场竞争模型来分析，组织的盈利与个人的投资理财便是财务与经济学常识的用武之地，组织产品的销售与个人 IP 的塑造都离不开营销与策划的技巧。当一切都用商业模型来看待时，便是产品经理真正的泛化。

1.3 互联网产品经理的规定动作

观察身边的互联网产品经理每天都在做什么，会发现他们使用频率较高的几个应用中一定有 Word、Excel、MindManager（或 XMind）与 Axure，于是不难得出结论：他们使用这些软件所做的事情莫过于需求调研、竞品分析与原型设计。关于这三件事，大多数关于产品经理的读物与书籍都有很详细的介绍，这里不标新立异，仅做简要介绍。

1.3.1 需求调研

360 公司董事长周鸿祎用六字真言形容互联网好产品：刚需、痛点、高频。说的就是用户需求对于产品规划的重要性。现代营销学奠基人之一，《哈佛商业评论》前主编西奥多·莱维特（Theodore Levitt）曾经说过的一句话也经常被引用以提醒产品经理们聚焦用户需求，这句话是“人们不想买一个 1/4 英寸的钻孔机，他们想买的是一个 1/4 英寸的洞”。

很多人也会列举亨利·福特的一句话——如果我问人们想要什么，他们一定会说要更快的马——来证明以前不存在的需求可以经过他们的双手被创造出来。这样的人也会相信乔布斯说的那句：“消费者并不知道需要什么，直到我们拿出自己的产品，他们就发现，这是我要的东西。”

表面上看似矛盾的两派观点，实则在于用户需求的底层却惊人地一致。它们都建立在理解用户深层次需求的基础上，并在不同行业使用不同的路径实现。正如营销管理理论中对产品层次的五层分级：核心利益、基本产品、期望产品、附加

产品以及潜在产品，在满足用户需求的核心利益的前提下，有的产品满足基本产品层次即可，有的会给出差异化的期望产品与附加产品，还有的产品则做到了潜在产品的开发。正因为满足了不同人群的不同需求，产品才丰富多彩，否则也不会有快捷酒店、轻奢酒店以及豪华酒店的区别了。

陈峻锐所著的《匹配度》一书主要介绍了用户需求调研的实务性，书中的两个主体部分分别介绍了需求的“采集”与“分析”这两个主要过程。

需求采集

这个阶段主要是“备”“找”“谈”三个环节。准备环节主要是准备好问题以及以大规模的信息搜集工作为代表的案头研究。在找寻访谈用户环节，需要找到典型用户与特例用户两类人，就如同我们衡量一个样本群体时会关注均值（bias）与方差（variance）一样，典型用户就是均值，而特例用户则是方差。访谈环节则是要用好嘴巴与眼睛这两个器官。当使用嘴巴提问的时候，要问满足开放状态的问题而非具有诱导性的封闭问题，问题也应该具体且明确，少抽象与含糊。而当使用眼睛进行观察时，则需要利用 AEIOU（活动、环境、互动、物品、用户）五个维度的原则来对观察到的信息进行分类。

需求分析

我们感到困惑往往因为两种情况，一种是信息少，一种是信息多。需求分析则是后者带来的困惑。当需求采集回来之后，需要进行分析与筛选，最重要的是从需求中找到进行产品研发的机会，为市场服务，这也算是一种以终为始了吧。

首先，在分析的过程中需要用到一些工具，无论是表格还是图形，都是依据一定的原则组织需求信息的形式。工具说到底是一些方法论与逻辑思考的模式，如矩阵分析（又被称为象限分析）、杜邦分析（结构或者层次化分析）、流程图、归纳与演绎。其次，整理好的信息需要被理解或洞察，洞察的过程往往因人而异，一些如逻辑推演、同理心带入等方式都可以使用。最后，就是为需求分析的结果付诸实践，类似沙盘推演等方式可以用来进行筛选后需求的检验，同样地，类似 ABTest 等直接投入市场的策略往往也可以起到检验的效果。

苏杰在《人人都是产品经理》一书中也针对需求谈了他的看法，并把调查问卷、用户访谈、数据分析与可用性测试分别在说与做、定性与定量的二维象限中

进行归类。除此之外，当需求调研与公司中的工作流程相结合的时候，还会遇到需求评审、需求冻结、需求变更等实际问题，这些都是产品经理需要面对的。

1.3.2 竞品分析

竞品是商业中的同类竞争品种，多种竞争并存时，则衍生出“战略”这种“高大上”的词汇。很多现代商业竞争往往向历史战争借鉴经验，从而衍生出商战的概念。老祖宗们早就为我们留下了一笔可观的财富，即中国最早的兵书，也是世界上最早的军事著作《孙子兵法》。网上有许多关于该书的解读，上海华与华营销咨询公司董事长，“超级符号就是超级创意”的华与华方法创始人华杉的白话解读版本《华杉讲透<孙子兵法>》就是一本将孙子兵法与商业相结合的著作。与之对应的卡尔·冯·克劳塞维茨所著的《战争论》则是西方近代军事理论的一座丰碑。

《孙子兵法·计篇》中提到：“夫未战而庙算胜者，得算多也，未战而庙算不胜者，得算少也。多算胜，少算不胜，而况于无算乎。”这句话强调了在事前了解竞争者实力并进行分析的重要性，分析越多，了解越多，胜算越大。《战争论》中提出的“主力会战”则强调集中优势打击敌人重心，似乎更像是在细分市场上夺得主动权，以获得利基市场。由此可见，在正式开工之前进行竞争分析或竞品分析十分必要。

竞品分析大致可以分为逐步聚焦的三个层次，分别是市场的角度，从企业的角度，以及从产品的角度。

市场的角度

既然是分析竞品，自然是自家的产品与友商的产品在同一片商业天空之下进行比较，因而从市场的角度分析竞品就意味着需要了解产品所打入的细分领域。重点来说是两方面，一是大环境，二是大用户。首先，了解整个市场的容量，对自己计划发展产品的“天花板”有一个了解。很多时候，这样的市场容量数据也可以在行业报告中获得。其次，针对市场中产品的关系人群进行分析，通过如地理位置、文化程度等人口统计学属性可以对人群市场进行细分，并逐个研究其付费可能、意愿与动机、使用场景等。例如教育行业的关系人群分为区域主管部门、校级领导、教师、学生、家长等几类，每一类人的付费可能性与使用产品的场景

各不相同，需要区别对待。

企业的角度

商海用来比喻商业的市场大环境，于是才有了激烈竞争的红海与鲜有涉足的蓝海之说。如果海对应市场，那么企业便是海洋中的生物，资本市场每天发生的融资、并购与倒闭则是海洋世界中的生长繁衍、弱肉强食与自然衰亡。生物界的法则同样适用于商业，了解同场竞技的选手有助于我们求生。对于企业来说，从公开的信息到小道消息都值得关注。

首先是公开的信息，企业的资本实力、子母公司、股东背景等都应该有所了解，如企查查、天眼查、启信宝等软件都可以将这些信息一网打尽。如果是上市公司，还应该关注其公开财报信息。其次是小道消息，通过特殊渠道可以了解企业的战略规划、商业模式、团队规模、运营现状、渠道代理等方面的信息。小道消息可以依靠内部人员披露、第三方机构调查以及非正式场合“道听途说”等方式获得。在对竞争企业建立多维度分析的过程中，需要综合利用公开渠道和小道消息。往往认为公开渠道信息较为可靠，但其实也不可全信，很多来自企业自身的公开信息往往带有自证式的广告宣传性质，遇到这种类型的公开信息需要取其精华，去其糟粕。一些小道消息虽然是道听途说，但是当结合上特定的背景，就变得鲜活又真实，可以帮助进行综合判断。

产品的角度

企业的产品是真正参与竞争的实物，是真正的竞品。大多数的竞品分析采用棋盘的形式进行，横向列出分析的维度，纵向给出待分析的竞品。按照实际情况将竞品在各个维度上的信息填入棋盘中，用作分析的维度包括产品的界面功能、业务流程、细分市场、收费标准，除此之外还可能包括产品中的数据与内容等情况。产品棋盘的分析方式重点在于比较，目的在于明确下一步的行动计划。在一个没有那么新的领域中，往往最快速的产品构建方式就是“抄”，然后进行微创新。因此，通过纵向比较，可以细致了解各个竞品在同一个功能点上的细节，并在我们为自己的产品设计这个功能时提供借鉴。

如果把一个产品团队看作一个小型的创业公司，那么竞品分析就是一次经营战略的制定。于是 SWOT 分析、PEST 分析、杜邦分析、波特五力分析、波士顿矩阵分析、KANO 分析等便都可以派上用场了。

竞品分析是一个细致而烦琐的工作，但是“磨刀不误砍柴工”，做好准备，才能有的放矢，最终在产品规划与设计上赢得主动与先机。

1.3.3 原型设计

设计是艺术的范畴，艺术又可以大致分为视觉艺术和听觉艺术，互联网中的原型设计就属于前者。在人类历史发展的长河中，从茹毛饮血到刀耕火种，从精耕细作到工业革命，视觉艺术也从旧石器时期的平面石刻演变成了栩栩如生的立体壁画，从文艺复兴时代后兴起的巴洛克与洛可可建筑风格演变成了模块化、扁平化的当代互联网设计风格。

在互联网产品研发过程中，原型设计的目的是为了给真实的产品开发提供参考，而产品最终是要面向用户的，所以不可避免地会把原型设计与用户体验联系在一起。

网页可用性行业大佬，《用眼动追踪提升网站可用性》的作者雅各布·尼尔森（Jakob Nielsen）提出的用户体验的“LEMErS”原则被业内奉为圭臬。这几个字母分别是易学（Learnability）、高效（Efficiency）、易记（Memorability）、纠错（Error）与满意度（Satisfaction）的缩写。卢克·米勒（Luke Miller）在其著作《用户体验方法论》中也推荐了这个方法。个中内涵与风靡全球的设计书《别让我思考》有异曲同工之妙。

易学

面对产品能够快速上手是很重要的，否则用户就会因为失去耐心而流失。郝志中老师在其所著的《用户力》一书中也给出了用户在面对产品时因为看不懂、找不到、选择多而产生想法从而影响产品效果的实例。

易记

如果易学是首次接触产品的用户的感受，那么易记就应该是与产品再次触电时用户的感受。这里的易记包含两个层面的含义：一是产品应该是渐变的，而不能让用户在前后两次使用时有恍如隔世的感觉；二是利用产品养成用户的习惯。

高效

产品是为了帮助用户达成某种目的，完成某项需求而产生的，它应该直截了当地解决问题，而不是兜圈子。谷歌在设计搜索引擎时提出的“用完即走”的主张便是高效最好的代名词。

纠错

用户难免在使用产品时有些跳脱，如何让产品在用户出错的时候还能够友好地进行展示，这体现了产品的鲁棒性，也对产品的细节提出了很高的要求。

满意度

用户的满意度主要分为两个层次：一是不强迫其做不必要的事，例如不必要注册的产品就不要为了收集数据而强制要求注册；二是给用户更多惊喜，例如产品中的彩蛋或者用户情感上的关怀与陪伴。

说完了原型设计的“道”，我们来关注一下“术”的层面。

工欲善其事，必先利其器。这里为大家介绍几款可以用得上的产品原型制作工具。Axure 已经是产品经理的必知必会了，它是一个介于线框草图和高保真原型图之间的产品原型制作工具。除此之外还有线框原型的绘制工具，如 Mockups。而 POP (Prototype On Paper)、InVision 与 Flinto 则适合用来完成移动端的页面链接交互。一些高保真的原型图则可以使用 OminiGraffle 、Sketch 或 Adobe Illustrator 来完成，只不过这部分工作往往是 UI 设计师的职责，除此之外的交互性原型设计则需要借助 Principle 的力量来完成。

工具不在于多，而在于精。除掌握工具之外，还要积累一些基本的设计思路，这样才能够完成一个原型的设计。闫荣老师在《产品心经》一书中给出了移动设计中关于导航、表单、表格、搜索、排序、过滤、工具栏、邀请、反馈、功能可见性、弹窗等十余种功能的总结，很是受用。作为一名产品经理，平时在试用产品时应该多做积累，建立自己的原型设计素材库。

第 2 章

产品经理的今生

- 2.1 卖家秀：自我提升的几项技能 20
 - 2.1.1 从需求文档到动机文档..... 20
 - 2.1.2 从竞品分析到广义竞品分析..... 22
 - 2.1.3 从原型设计到交互设计..... 24
- 2.2 买家秀：弄垮团队的若干“要领” 28
 - 2.2.1 越过产品雷池 28
 - 2.2.2 踏入团队雷池 29
 - 2.2.3 迈向公司雷池 30

产品经理的前世打下了一个好基础，今生发展如何就要看当下的修为了。为了自我的提升，产品经理需要在前世的规定动作上再做技能升级，方得圆满。很多产品经理的培训课程为大家提供了产品经理的职业技能规划，为从事这个职业的产品经理们描述了一幅广阔的未来前景，这算是一种“卖家秀”。而实际工作中，大家遇到的问题却千差万别，往往记不起理论，或是照搬照套，因而作茧自缚，成了产品经理的“买家秀”。

2.1 卖家秀：自我提升的几项技能

之前提到的需求分析、竞品分析与原型设计中的内容是产品经理必知必会的，由于上述工作大多在纸面上完成，所以应该算是 paper work，即纸面上的功夫。其背后深藏着的思考，才值得产品经理咂摸。

2.1.1 从需求文档到动机文档

需求说明文档是产品经理经过需求采集、分析、筛选、提炼而最终形成的用于指导产品研发的文书。经历过需求分析的产品经理都会有很清楚的感受，往往分析得出的用户需求真正付诸于开发实践后，用户并不买账。

疑惑随之而来，难道每一次的分析不是按照需求分析的要领进行的吗？为什么进行需求评审的时候大家一致认为可以投入研发的需求，反映到最终产品上却是门可罗雀？这里主要有两个问题，一方面是在进行需求分析和评审的时候容易犯“自证式谬误”，即从方法论的角度很容易给每个自己想要支持的需求找到理论依据。举一个最为常见的例子，在微信朋友圈中，我们往往会选择去看那些观点被我们支持的文章，而对一些反对的声音嗤之以鼻。因而当一个观点被重新提出的时候，若与自己的观点相同，便听之任之；如果与自己意见相左，则很容易找到证据反驳。因此，一个需求是否迫切且真实，不在于能否找到证据证明它，而在于是否有足够多的声音反驳它。另一方面，由经验累积而成的实务性方法论往往总结自成功的案例，且这些案例的当事人也是较为活跃、乐于分享的人，这里本身就存在“有偏采样”。怎么理解呢？首先，失败案例是大多数，而我们只对成

功的案例进行采样,即便失败的案例使用同样的实务性方法论,可是它却沉默了。其次,分享者毕竟是少数,没有分享却使用了完全相反的方法取得了成功的案例,对我们也是不可见的。因而,亦步亦趋并不一定能保证获得相同的结果,更何况,影响结果的因素多种多样,真实场景下难以实现变量的控制。

撇开方法论不谈,有一点是产品经理公认的真理——需求和需要不是一回事。也就是说,听到的不一定是真的,更需要看到,看到的也不一定是真的,还要了解用户是怎么想的。需求一词已经泛滥,作为一名产品经理,当有人和我们谈需求的时候,要立马警戒起来,要问自己,这到底是什么“需求”。在大多数情况下,别人和你说的需求不过是“想要”罢了,其背后的动机才是需求。那么要怎么在众多需求中找到最有价值的那一个呢?这里介绍三个小方法。

需求倒排序

李开复老师曾经分享过一个“报纸头条测试法”,大体意思是,如果有两件事同时被搞砸,并且会被刊登在明天的报纸头条上,哪一件会让你觉得更加羞愧,那么这件事就是对你最为重要的事。这种方法很适合用来做出遵从内心的选择,也可以使用在需求的选择上,还可以帮助有选择和排序障碍的人快速做出抉择。

这样的测试方式在心理学中也能见到——“家庭船”亲子实验,适用于一些纨绔少年或者叛逆子弟。实验是这样的,让叛逆的孩子与父母以及爷爷、奶奶、外公、外婆一起待在一艘虚拟的船上,经过主持人的情境引导,需要孩子在每一次“家庭船”遇到艰难的时候,把一个人丢下船,只有这样全家人才能生存。随着实验的进行,孩子受到的道德谴责越来越大,因而反思也越来越多。大多数的叛逆孩子最终在实验结束时都会和父母家人相拥而泣,可见其效果卓著。

以上两个思想实验都运用了逆向思维,当我们在所有需求中难以抉择哪个更重要的时候,不妨反过来进行倒排序,即如果要去掉一个需求,应该把哪个去掉,这样最终留下来的就是最有价值的需求。

马斯洛层次需求理论

马斯洛自己可能也没有想到,在多年之后,会有一群互联网行业的从业人士成为其拥护者。马斯洛的层次需求将人的需求分为三个大层次及其包含的五个小层次,即生存(生理需要、安全需要)、归属(社会需要、尊重需要)与成长(自

我实现需要)。越是底层的需要就越刚性，越有广泛的市场。还有一种需求分析的声音，是从佛家的“贪嗔痴慢疑”角度入手来进行筛选的，这也是一种思路。总之在评估用户需求的时候，不妨对应层次需求进行理解，或许能够有所启发。

丰田五问模式

这个方法有一个更容易被人理解的名字，叫作“打破砂锅问到底”，其发明者是丰田汽车公司创始人丰田喜一郎的父亲，丰田佐吉。丰田佐吉在经营丰田织机这家企业的时候发明了这种提问方式，即针对一个问题采用连续追问五次的方式找到本质原因。例如，大数据时代很流行为企业引入商业智能（BI）系统，关于企业为什么需要这个系统，我们就可以采用连续追问的方式找到原因。为什么企业需要商业智能系统？因为企业需要了解信息。为什么需要了解信息？因为需要对各职能部门有把控感。为什么需要有把控感？因为害怕企业运行出错。为什么怕出错？因为出错会影响企业业绩。为什么怕影响业绩？因为对自己的职位有影响。分析到最后，其实是因为企业主缺失安全感，为了满足这一点，就可以实现与其他企业的商业智能系统不同的功能。

2.1.2 从竞品分析到广义竞品分析

传统意义上的竞品是在同一个商业市场下的竞争性品种，在很多情况下，这样的竞品应该属于同一个分类。粗糙来分，电商网站间是竞品，如聚美优品、唯品会与天猫商城等。精细来分，可以在更为细分的市场中寻找竞品，如专注于母婴市场的贝贝网、苏宁红孩子、万达孩子王等。

分析竞品最重要的是有条理，有详情，有结论，对应于竞品分析的规定动作。这里再介绍一种移动产品竞品分析的方式，即从空间和时间维度来进行分析。

空间维度

空间包含了移动应用产品本身，以及该产品的周边配套设备。研究产品本身可以从竞品的交互体验与用户流程上进行多维度的剖析。产品的周边配套设备则涵盖了经营战略与商业模式等业务分析。

时间维度

在时间维度上，我们可以针对移动产品的历史版本进行剖析。许多网站可以查看移动应用历史版本，如手机中国（app.cnmo.com）等。在这些网站上，我们可以检索一个应用的历史版本，查看每个历史版本中更新了哪些功能，这有助于刚进入市场的产品团队抓住核心功能进行排期研发，而不是一开始就将竞品的所有已知功能等同视之。除此之外，从历史版本的描述与截图中，我们还能找到一些产品运营策略的蛛丝马迹，这些都是可以借鉴的。每一个竞品的历史版本，有可能就是当下产品的未来版本。与其自己尝试，不如对竞品的经验运用拿来主义。

在传统的竞品分析之上，还有更为广义的竞品。看待市场的眼光与胸怀决定了我们的竞争对手和未来的前途。这里介绍几个关于广义竞品的观点。

潜在竞争

传统竞争中的竞品是直接能看得见的，在波特五力模型中，这样的竞争者被称为“行业内对手的竞争”，除此之外，还有两种竞争类型，一种是潜在进入者，另一种是替代品。潜在进入者指的是和我们产品类型、业务结构相似，技术积累雷同的有潜力的竞争对手，它们只不过暂时还没有涉足我们所在的市场。而替代品则可以使用一个例子加以诠释，众所周知，可口可乐与百事可乐互为竞品，脉动与激活互为竞品，它们各自的市场分别是碳酸饮料与维生素饮料，但它们本质上都是满足人解渴这一需求，在这个需求点上，矿泉水就是上述饮料的替代品。无论一个人多忠实于一个品牌，当遇到极限条件时，如口渴时买不到任何饮料，也不得不选择其替代品。

作为某产品的潜在进入者，当需要涉足该产品所在的市场时，至少应该回答三个问题：为什么要做这个产品（目的）？做这件事有什么优势（装备）？这件事要怎么做（路径）？有了目标，带上装备，看到路径，才能在“打怪升级”的路上越走越远。反之，作为某产品的负责人，在考虑潜在进入者的时候，也要问清楚自己三个问题：对手为什么没有做？在对手做之前，我该做些什么？当对手开始进入的时候，我该怎么办？针对这些问题，波特给出了成本领先、差异化、集中等战略。这些问题的解答归属于经营战略，展开略微复杂，又与本书主题相背离，故而略去。

跨界竞争

跨界是近些年的热门词汇，以前井水不犯河水的两个行业，如今由于跨界而产生竞争。歌手兼职演员是跨界，互联网企业投资拍电影是跨界，自媒体涉足电商也是跨界。以前的行业分割在当下变得越来越模糊，而跨界之所以能够发生是基于所跨的两个行业间的共同点，或是用户基础，或是共享技能。

我们通常进行的竞品分析看似有宽度、有深度，但如果放在整个产业或商业大环境下，其实是一个细而窄的领域。这个行业与社会上的千万种业态共享着劳动者的智力、消费者的钱包，因而也具有可以相互借鉴与学习的意义。论文案，快消品领域不输互联网；评创意，广告公司自然拔得头筹；讲服务，餐饮业率先垂范。而写文案、想创意与提供服务等技能又是互联网产品经理必须具备的，因而留心生活中的点滴，用跨界思维向非竞品学习，这是产品经理的高标准要求。

时间竞争

罗振宇老师在 2017 年跨年演讲中提到五个主要观点，“时间战场”位列第一。其主要的论断是，在面向消费者的行业中，时间资源成为了刚性约束，无论做直播、社交、电商、餐饮、游戏还是旅行休闲，看似不相同的行业实际上都在竞争用户的时间。因此一切皆竞品。

在可预见的未来，如何在消费者的有限时间中抢占一席之地则代表了产品的黏性。从这个角度看来，音乐、电台、类似《吴晓波频道》与《得到》等知识付费节目成为了竞品，它们抢占的是开车族的上下班时间。餐饮、电玩、影院等休闲场所也成为了竞品，它们抢夺的则是休闲的时间。这一再提醒我们，形成竞争的不再只是功能相似的产品，把产品还原到用户使用的场景中，找到在那个场景中的替代者，它们都是时间上的竞品。

2.1.3 从原型设计到交互设计

处女座的人往往会被认为是从事这个行业最适宜的人群，原因之一就是其在原型设计的时候会更加苛刻、追求完美。然而进行原型设计并非产品经理的终极技能及长远目标，互联网产品领域乃至更为广泛的工业产品领域的产品设计风格和原则才是需要关注的。

设计被认为是产品研发的前序步骤，好的设计能够引导好的研发过程，进而促成好的市场反应，而不好的设计则常常给我们带来困惑，生活中屡见不鲜。相信下面这些场景你一定亲身经历过。

你是否在面对一扇门的时候不知道是该推它还是拉它，以至于去拉一扇本该被推的门，去推一扇本该被拉的门，而对于一扇左右滑动的门无论推拉都无计可施？你是否在使用了带有蓄水功能的洗手池后不知道如何将水池中的水放掉，以至于最后将手插进已经被污染的水中，向下按动底座，使污水流走？在洗手这件事情上，我有亲身体会，某个医院的洗手池上明明贴着“请按动出水”的说明，而水龙头却是旋钮式的。诸如此类的还有仅仅备有一卷卷纸的卫生间。很难想象卷纸尚未用完，但似乎又不足以在下次补充前满足客户需求，保洁人员究竟是应该用全新的卷纸更换它使之浪费，还是应该让这段时间内的客户无纸可用而窘迫？



图 2-1 令人困惑的水龙头

除此之外的不好设计还包括，具备超多按键的遥控器，家中电器的总控面板，无人引导的现场购票队伍，以及一些装配豪华屏幕的廉价汽车。

如果你对上述这些设计带来的困惑感同身受，或许你就可以理解如何从原型设计过渡到交互设计了。

通常意义上，设计可以分为三种类型，分别是工业设计、交互设计以及体验设计。三者关注的侧重点各不相同，但共享一些总体的设计原则。

工业设计

素有设计领域“奥斯卡”之称的德国红点设计大奖（Red dot design award）是工业设计领域人员的追求。在手机和生活用品领域，中国的互联网公司曾多次获奖。在工业设计方面，设计师更加关注外形与材料。从塑料到金属再到陶瓷，这是一种材料上的升级，从有边框到窄边框再到曲面无边框，这是外形上的升级，手机就是靠这些方面的工业设计一步步进化的。

交互设计

交互设计关注的是使用者和产品之间的互动情况，当产品变成互联网产品或者以计算机为载体的产品时，交互设计退化为人机交互（Human Computer Interaction, HCI）。交互设计领域崇尚的是易懂性与易用性。为传统的工业制成品空调所配备的遥控器就是典型的交互设计的例子：上面的每个按钮是否都会被用到？用户是否理解都是什么意思？一个简单的用于评价交互设计好坏的标准就是用户是否需要说明书才能进行操作。在人机对话方面，从关键词查询，到选定问题回答，到自然语言的开放式回答，这些都是交互设计的进步。

体验设计

体验设计更关注用户的情感。这里的情感可以包含视觉上的审美、触觉上的愉悦、听觉上的轻松以及心理上的舒适等。产品用色讲究审美，红色和橘色给人积极向上的感觉，蓝色给人冷静的感觉，黑色给人消沉与酷的感觉，这是颜色传达给用户的视觉感受，也能够影响人们的情感。一本书的装帧设计能够给读书人在尚未翻开书前埋一个伏笔，封面是烫金还是磨砂，纸张是铜板还是胶版，给人的感觉完全不同。电话接线员查询信息时的等待音是嘟嘟声还是莫扎特的曲子也会影响客户对客服的评分。排队等待时的安排也能体现商家在安抚人心上所下的功夫。

作为全球最具影响力的设计师之一的唐纳德·A·诺曼（Donald A. Norman）博士曾著有《设计心理学》系列书籍。在书中，诺曼博士给出了设计的七个基本原则。

示能

示能就是事物示意用户能用来干什么，这是事物的固有属性。当一个产品摆在我们面前时，大到桥梁和马路，小到指甲刀，我们都可以通过多种感官去了解这个产品的属性，这算是认知产品的第一步。

意符

意符是示意的符号，好比桌面上的便签纸，可以提醒我们某个物品具备什么样的属性。例如马路上的箭头指示我们行车的方向，箭头就是意符；网页上的登录按钮或引导标志指示下一步操作，这个按钮就是意符。

约束

约束则是事物叠加了现实因素的情况。VR 无论如何逼真也无法实现真的触碰，

这是物理约束；互联网页面两侧频繁闪动的裸露图片可能会带你前往一个“危险”的页面，这是文化约束；网页上的灰色字体表示不可点击，这是语义约束；验证码输入框旁边的图片看不清的时候，人们知道要主动点击以获得新的验证码，这是逻辑约束。

映射

以“相似相邻”为主的格式塔原则是映射最重要的部分，其意义是，具备相同功能、反映相似含义、携带近似属性的事物应该在逻辑位置上和物理位置上接近。

反馈

正如我们在前面所讨论的那样，控制论中强调闭环，管理学中强调沟通，这些都是反馈的实例。在设计中给出反馈仿佛产品具备人性一样能够与人沟通，在输入错误密码时，页面通过文字提示或输入框仿佛摇头一般左右震动提示不正确，这都是反馈的典型代表。

概念模型

概念模型就是人们对事物认知的模型，简单来说就是人们用来解释某个功能背后原理的一套说辞。例如冰箱制冷的原理和空调制冷的原理略有不同，空调制冷在大多数人的印象中是靠送冷风实现的，我们称之为风冷式，因此送风制冷就是一种概念模型；而冰箱制冷除了风冷式之外，还可以通过冷凝液蒸发（依靠蒸发器）直接进行制冷，我们称之为直冷式，直冷式又是另外一种概念模型，这个模型相对风冷式来说更难以理解。

可视性

无论是意符还是约束，都需要对用户保持可视性。直观可见可以保证易懂性与易用性，因此也是好的设计原则。换句话说，好的产品自己会说话，能让用户一看就懂，而不需要配备一个解说员。

由此可见，产品设计远不止原型设计那么简单，从用户的角度出发，以人为本，方能赢得市场和人心。

2.2 买家秀：搞垮团队的若干“要领”

如果把阅读看成知识的输入，那么实践检验以及反思写作便是一种输出，通过输出看到成效并与期望产生对比才能知道输入的正误，便于有的放矢地针对性输入，这与之前提到的闭环思路如出一辙。

正如“听过了许多道理，却依然过不好这一生”这句话告诉我们的，“听”道理的输入与“过”一生的输出并不一定呈正相关。因此产品经理知道许多方法论与技巧，也不一定与工作可以得心应手成正比。

我把引导行为的道理大致分为两类，一类是旗帜型，一类是雷池型。所谓旗帜型，就是列举正面行为的道理，大多数的方法论、成功学属于此类，它们给出的往往是条件或过程，听得多了容易让人将其与结果产生必然的因果联系。而实际情况是，不同的方法论有各自适用的场景，共性的方法论又太过粗粒度难以实践，最终造成学习完成后倍感充实，回头细想却又怅然若失的感觉。另一种雷池型，则属于逆向行为的总结，即枚举不要做的行为，是容易导致坏结果或产生消极影响的行为。这种从反向来思考的方式，可以给出一个清晰的列表进行自查自纠，虽然不是立竿见影的大力丸，却是能够防止腐坏的预防剂。

产品腐坏，无非就是产品涉及的各个干系部分出了问题，没有运行在其正常轨道上和位置中，这个不当的轨道和位置就是雷池。据考证，古雷池是一片位于长江北岸安徽省望江县、宿松县以及湖北省黄梅县一带的水域。《报温峤书》中写道：“吾忧西陲，过于历阳，足下无过雷池一步也。”意思是让温峤坐镇原防，不要越过雷池向东行军，后来用于表示不可逾越的一定范围。

下面就来看看，作为产品经理，我们都涉足了哪些雷池。

2.2.1 越过产品雷池

产品是产品经理最直接可操纵的，要想“使坏”，可以恣意妄为。

做多不做少

在规划产品第一版本功能的时候，把产品功能规划得尽可能多，只有多才

能显示出考虑全面。所有的需求都是我的孩子，我舍不得扔掉任何一个，不抛弃、不放弃任何一个需求，“一个都不能少”啊。搞跨产品的第一要素就是越复杂越好。

我思故我在

对 UI 说：“这个功能设计之初不是这个样子的，你用的这两个颜色配起来有点怪怪的。”对开发说：“你难道就不会使用那个团队的技术吗？他们都可以实现，为什么我们不可以？”对自己说：“我是乔布斯、张小龙一样的产品经理，只有独树一帜，才能体现出实力。”处处以自己意志为标准，其余一切都是浮云，忽略集体智慧，是搞垮产品的又一必杀技。

需求变变变

有的时候产品经理还会制造一些特殊情况，例如临时取消某需求，又如临时增加某需求，再如改变后要求将需求改回从前的状态，开发人员会非常无奈。经常性变更需求，产品久久得不到交付，可能看不到产品最后的样子，它就已经胎死腹中。

功能抄抄抄

张三家的这个功能不错，我们做一个。李四家的这个效果酷炫，我们也做一个。王五家的配色很好，我们模仿一下。恭喜你，你已经学会了如何做一个抽象派的产品经理，毫无业务逻辑可言，而只是拼凑。

2.2.2 踏入团队雷池

要想做一个“合格”的猪队友，别忘了给团队挖点坑，下面是一些挖坑的要点，请收好。

黑锅大赠送

产品做出来根本不是我原来规划的那样，开发这个实现不了，那个来不及做，开发被黑锅砸中一次。这个 BUG 我早就提过了，开发怎么还不修改，开发被黑锅砸中两次。总之，不按时交付，是开发的问题；产品不稳定，是测试的问题；业绩不景气，是运营的问题。出事第一时间撇清关系，团队成员会瞬间反感。

老板大于天

领导需要某个功能，所以我们要做。领导一周后有个接待，我们临时做一下。领导不喜欢这个，我们得改。凡是领导说的，不容置疑都是对的；凡是领导说的，都要不假思索地落地。不加以考虑地服从上级有可能会连累团队的兄弟。

同学快干活

在产品经理看来，开发人员只要把手头上的事情做了就行了，而且他们提的想法也不现实。若开发人员阅读课外书籍，产品经理的内心独白是：“你看这些书干什么，它们能帮助你写代码吗？”这种只问目标结果，不问个人成长的沟通方式，很快就會把团队的积极性抹掉。

华丽传话筒

领导说要做这个，肯定有他的道理，我们只管做就是，不要问为什么。这个需求已经传递给大家了，大家清楚了就可以行动了。不介绍背景、不同步目标的传递方式会让团队快速“沉浸”在细节工作中而难以有全局观，这样统合起来的工作难以形成真正的合力，发挥最大价值。

2.2.3 迈向公司雷池

不满足于搞垮团队？作为高阶的产品经理，再来给公司补一脚。“爱之深，恨之切”用来形容这种心情最恰当不过了吧。

管理我最爱

各个团队请把这个表填一下；请各位产品经理报送一下自己的产品现阶段的情况；请大家填写一下绩效；请大家做一下周边同事和领导的评价表格；周报发一下……每周都有新表格，每天都有新感觉，这种过度的管理，很容易把公司员工弄得疲惫不堪。结果是周报越写越长，工作越做越少。

醉情 KPI

给研发定个 KPI 吧，每周代码量多少行。给测试定个 KPI 吧，一定要测试出多少个 BUG。给团队定个 KPI 吧，一定要支持多少个项目。给个人定个 KPI 吧，一定要做多少次分享。技术团队的 KPI，容易让大家陷入应付式的满足，不妨试

试 OKR。

画饼高高手

在向高层介绍某技术的未来前景与规划，给公司带来的价值，以及团队的应对方法时，每个季度换一种介绍方式和内容，但每一种都能够持续落地与执行。这种持续画饼的能力，不得不承认也是一种功夫，只是常立志，不如立长志。

外来和尚好

这个业务应该这么做，因为 BAT 也是这么做的。这个事情得找他来做，他是从 BAT 出来的，肯定了解方法。抛开行业属性做产品，一味追求跨行业的成功模式，只会让公司骄傲自大。

三省吾身，面向上述所列，也偶有犯之，愿与尔携行，有则改之，无则加勉。

第 3 章

产品经理的入行

- 3.1 入行做产品的几种可能 34
 - 3.1.1 源自技术岗 34
 - 3.1.2 源自业务岗 35
 - 3.1.3 源自应届生 36
- 3.2 上岗后的第一件事 37
 - 3.2.1 产品全图 38
 - 3.2.2 行业全图 39
 - 3.2.3 产业全图 40
- 3.3 工作中如何学习 41

都说产品经理是没有门槛的职业，意思是从事产品经理这个职业并不需要预先储备什么硬技能，或者说硬技能可以在短时间内习得。与之相对应的行业，如医学，则需要长期的预先教育与学习。技能越是轻描淡写，越是袖里玄机，越不可被量化与评测，也越是难以提升与进阶。

不知是什么样的魅力，引得无数英雄为产品经理岗位竞折腰。既来之，则安之，既然入了行，就要想办法 make a difference。英语中夸赞一个人杰出，用的是 outstanding，即 stand out（醒目）。不难理解，所有人坐着的时候，你站起来了，自然就醒目了。无论大家在产品经理这一行成绩如何，这里都有一些可以让你 stand out 的“银弹”奉上。

3.1 入行做产品的几种可能

行内有一句人尽皆知的话叫作“人人都是产品经理”，一方面是说产品经理门槛比较低，谁都可以张口对产品说上两句，品头论足一番，说出些似是而非的道理；另一方面，这句话也给那些喜爱产品，想转行做产品经理的人信心，使他们相信，只要坚定信念，成为产品经理并不困难。

成为产品经理的人往往有如下一些来源，我大致分成三类：一是来源于技术岗位，二是来源于业务岗位，三是来源于应届生。

3.1.1 源自技术岗

在企业中，从事技术岗位的有开发、测试、运维工程师等。如果算上数据分析师、数据挖掘工程师以及算法工程师等的话，种类就比较繁多了。技术岗位相较于业务岗位偏重于后端，即着重于怎么做，而不是为什么要做。

技术人员做产品经理有一些优势。首先是对于技术具有品鉴力。现在产品中的需求往往伴随着数据分析或者人工智能技术而被提出，从事技术岗位的同事有能力判断某个需求是否可靠，也能够结合技术的趋势为产品设计新颖的功能点。其次是对需求实现时间的评估更为准确。尽管项目整体的把控仍然是项目经理的事情，但有的时候，产品经理也会帮助项目经理进行一些项目的管理工作。作

为一名从技术岗转行的产品经理，能够在需求确定的基础上更为准确地估算开发周期，从而更好地控制产品开发周期。除此之外，源自技术岗的产品经理在与研发人员进行沟通的过程中更容易唤起同理心。网上盛传的技术鄙视产品经理不懂技术还乱指挥的例子在这种情况下几乎不会存在。

技术人员转做产品经理也存在着一些需要补足的潜在技能。一是对外的能力，技术岗位的同事往往终日与代码为伴，很少有外出见客户的机会，转岗成为产品经理之后，无论是用户调研，还是与其他部门合作，都需要对外沟通的能力。二是向上的能力，产品经理承担了组织产品需求评审等职责，还需要定期向上汇报产品的进展与市场反馈，这些工作的性质都是组织信息并向上汇报，学会向上管理并争取资源推动所在的项目也显得尤为重要。

我也见过一些产品经理希望转行做技术，他们的理由是，产品经理岗位没有“技术含量”（产品经理们勿怪）。一般来说这样的产品经理年资较浅，还未能领悟这个岗位的精髓与奥妙。从技术岗位转成产品经理的，原因各不相同，有的是在开发与测试过程中接触产品并觉得新奇好玩，想要跃跃欲试，这与本身性格有关，选择产品就意味着自愿放下修炼多年的技术，实属不易，毕竟在这个技术飞速变化的时代，放下后想要再捡起来就没有那么容易了。还有一些人看不到职位上升的通道而决定转岗，虽然产品经理并不是真正的经理，但是对于产品的掌控力还是能够给予他们无冕之王的感觉，就是有些苦涩罢了。也有一些女性开发人员转岗产品经理，这往往是出于照料家庭的考量，可以理解。在女性转岗产品经理后，往往会在沟通与亲和力上获得优势，如果又是技术岗位出身，则能为其后续的工作增色不少。

3.1.2 源自业务岗

业务岗位包含运营、销售、咨询等岗位。这些岗位从性质上来说属于前端的岗位，相较技术岗位来说，与客户打交道更多。

业务岗位转岗产品经理，优点有三。其一，对客户需求更敏感，理解更加到位。业务岗位的性质决定了他们每天的工作就是与客户或者准客户打交道，见的人多了，需求收集自然也多。产品经理的需求来源之一就是销售的反馈，从业务

岗转做产品经理之后，之前的客户关系往往还能够为其工作提供一些思路，建言献策，相当于刚入产品岗，就已经有了自己的一些天使用户。其二，市场意识更强烈。业务岗位人员对于成本的考虑会比技术人员更慎重，但凡涉及资金的流入/流出，他们都会考虑投入产出比。而且，由于在外奔波次数较多，竞争对手自然也不少见，对于竞争对手的分析必然也更为透彻。其三，沟通能力有保障。业务人员最不愁的就是见人与说话，自然练就了一副好的口才，这是作为产品经理的基础能力。不仅如此，由于习惯了被拒绝，心理承受能力也足够强，面对评审对于产品的批评也自然相较其他人有更强的抵抗力。

业务岗转做产品也有先天的不足，最为重要的就是对技术的生疏，可能会导致需求规划过于保守或过分超前。在确定需求的时候，往往需要与技术人员交流更多，从而获得口头的许可，了解所列需求是否靠谱。

在我的身边，从业务岗位转为产品经理的并不多。大多数的人因为排斥频繁出差而希望获得一份安定的工作，我表示理解。

3.1.3 源自应届生

应届生完全不同于工作了一段时间再转岗产品经理的人员。应届生像是一张白纸，选择做产品经理需要慎之又慎，最重要的是得对产品充满兴趣。苏杰在他的著作《人人都是产品经理（纪念版）》中给出了若干建议，即兴趣、定位以及相关工作。

在面试申报产品经理岗位的应届生时，我听许多面试官提到过他们的关注点，总结起来大致有三：其一是对这个岗位的认知，其二是相关经验，其三是软素质等。对于第一点，作为应届生不能连产品经理是做什么的都不知道，起码得知道产品经理的职责，并且阐述一些自己对产品经理岗位的理解。对于第二点，很难要求一名应届毕业生有相关的工作经验，从相关程序来说，在校期间有过产品经理岗位实习经历的为佳，其次是在校期间进行过大学生创业活动实践，其本质也是三五好友在一起做小产品的过程，这也可以等同视为产品经理的相关工作，不过要在面试时清晰阐明自己所承担的工作与内容。前面两者都不具备也没有关系，如果上过一些产品经理网站，自学过一些视频教程，能够运用一些产品经理的日

常工具，也算是满足第二点要求。对于第三点，软素质包含面较广，面试官更看重勤奋与踏实，脑子活自然很讨巧，不过能够收敛过分发散的想法则是更为难能可贵的品质。

我大学时的一位师弟，就是通过校招进入产品经理岗位的最好例证，他的准备或许可以给应届毕业生们一些启示。他将要毕业时，我与他交流，当得知他的求职意向是产品经理岗位时，我便很感兴趣，想了解他为了应聘做了哪些准备。他给我发了三封邮件，其中两封邮件的附件是六个关于在线教育行业的竞品调研 PPT，第三封邮件是一个阅读清单，其中分门别类地清楚罗列了他阅读的 50 本书，领域涵盖了与产品经理相关的产品概论、信息交互、营销战略、新思想、用研心理、项目团队、数据分析以及其他。那是 2013 年 8 月，后来他去了微信团队，成为一名产品经理。

做产品经理并不总是能享受产品成功时的喜悦，也并不一直都是兴趣驱动的无怨无悔，更不是所有人都是受乔布斯或者张小龙的影响而立志成为影响亿万用户的高阶产品经理。如果你虚怀若谷，无须多言，扬鞭自奋蹄便是你的常态。如果你并非志存高远，而只是将其作为一份工作，我也依然保持理解，至少可以保持一种学习的心态，在完成本职工作的基础上不给队友添麻烦。

3.2 上岗后的第一件事

有一句戏谑却又不失真实性的话叫作“产品经理是 CEO 的学前班”。之所以称之为戏谑，是因为并非所有的产品经理都能够成长为 CEO，而真实性则在于，每一位 CEO 很大概率上都拥有独到的产品观与自己的产品哲学。

成为 CEO 并非目的，但学着像 CEO 一样思考却很有必要。把这样的思考与产品工作结合在一起，或许能够让产品经理的道路顺畅许多。我个人倾向于使用全图的概念，有的时候也称为“一张图看懂 $\times \times \times$ ”。对于产品经理而言，在从事这个岗位的过程中应时刻握有三张全图，分别是产品全图、行业全图、产业全图，这样才能全局考虑，条分缕析。

3.2.1 产品全图

非创业公司内往往存在多条产品线，甚至是业务线。产品线是比业务线更为微观的一个概念，一个业务线往往包含了多条产品线，如一个大数据业务下可能分为中心化产品与交付型产品等，多个产品、业务之间共享人、财、物等资源，也在基础数据、技术架构、市场环境等方面有诸多共通点。

作为产品经理，尽管经常关注的是自己团队的一个产品或若干功能，但是养成眼观六路，耳听八方的习惯可以帮助自己树立全局观念。

制作产品全图应该包含但不仅限于以下内容。

组织架构

产品是团队工作的产物，因此产品离不开与之相关的人。按照人事组织架构梳理出公司内各个产品的负责人以及相关干系人，能够形成公司业务与产品的倒置树状结构。按图索骥，可以在与相关产品交互的时候快速找到重要干系人，针对其所处位置采取不同的方式方法进行沟通。除此之外，在工作的过程中会与公司内各种人一起开会，也自然会认识许多的同事，把这些人按照整理出的空间框架摆放就位，可以让工作信息尽在掌握，也可以避免出现遗忘某人的工作性质与职责等尴尬情境。

时序历程

公司的产品都有其发展历程，并非一开始就是现在的模样。过程中存在着迭代、合并、撤销，也存在着由于关键节点战略目的或纯财务目的而产生的看似鸡肋的功能或产品。在时间上了解事业部、业务线的来龙去脉，会缓解看不惯、不理解、瞧不上的心理。换位思考，如果把产品放到特定的场景下分析，未必当时就能够做出更为明智的决策。从时间上了解产品历程还可以帮助我们避免走回头路，特别是对半路接手的产品，如果不做历史了解，就会陷入产品回滚的怪圈。

功能交叠

这种情况往往出现在大公司内部，特别是 to B 项目中。多条业务线的产品往往会出现功能交叠的情况，即不同的产品中都包含同一个功能，并且还处于独立研发的状态。首先，在一定程度上，大公司内保持一定的冗余状态是可以接受的。

其次，要区分交叠功能在不同产品体系中的应用场景，知道具体情境下该找什么样的产品团队。

3.2.2 行业全图

行业在百度百科中的定义是，按生产同类产品或具有相同工艺过程或提供同类劳动服务划分的经济活动类别，譬如教育行业、移动互联网行业等。行业中不同企业之间基于相互需要与合作关系便形成了价值链、供需链等。

以教育行业为例，主体是面向学前、K12 以及高等、职业市场提供教与学服务，其上游可以是支持教辅内容、教育装备以及信息化建设的企业，而其下游则可以是负责就业、人才招聘、留学与移民等公司，周边还可以包括金融、置业、消费与旅游等行业。这些行业与企业总体构成了教育行业的结构。

以自己的产品为中心，找到行业中生产类似产品的企业，进行进一步分析，可以尝试采用我总结的“5D”分析模型，即从人（用户群体）、财（商业模式）、物（产品形态）、技（核心技术）、境（使用场景）五个维度对友商进行分析。

用户群体

对用户群体进行划分，往往是根据人口统计学属性进行的，如年龄、性别、工作性质、兴趣等。仍以教育行业为例，以年龄为主，以兴趣为辅可以划分为学前、小学、初中、高中、大学、职场教育、兴趣培训、综合培训等若干类。从学前到综合培训，大致可以认为用户的年龄是逐渐增长的。在分类的过程中，往往难以满足 MECE 原则，如兴趣培训可以穿插于各个年龄阶段，可以不用较真，允许有一定的重叠。

商业模式

“刀片+刀架”是一种商业模式，讲的是售卖剃须刀的时候，为利润较低的刀架定较低的售价，而为利润较高的刀片定较高的售价以谋求盈利的模式，现在很多软件售卖也采用基础软件免费，按服务收费的模式。这里涉及的商业模式的是一些业务的组织形式，例如教育行业大体可以分为两大类，即 O2O（线上与线下），很多产品是跨两端的。线上培训的模式又可以分为 B 与 C 主导的模式，包括 B2C、B2B2C、C2B、C2C 等。

产品形态

产品的定位决定了产品的形态——做工具还是做内容，做社区还是做服务，做平台还是做交易中介。凯鹏华盈（KPCB）每年发布的《互联网报告》备受瞩目，其作者，有互联网女皇之称的玛丽·米克尔（Mary Meeker）在报告中给出了一个概念：Internet Trifecta。Trifecta 原本指的是赌马时一举猜中前三匹马的连下三成，用在互联网领域表示内容（Content）、社区（Community）以及商业服务（Commerce）三大业务与产品形态。

核心技术

各个行业使用的核心技术并不相同，这些技术往往可以通过一些行业报告、公司内的技术专家或者技术极客了解到。以教育行业为例，一般会用到因特网技术、人机交互技术（如触摸屏、体感、手势等）、人工智能技术、沉浸体验技术（如AR、VR、Game）等。

使用场景

关于使用场景的分析因人而异，产品经理可以根据行业特色进行逐一归纳。依然以教育行业为例，可以把学习分为定时定点与不定时不定点，而接触学习的手段也从线下转移到了 Mobile、PC 和 TV 上。

3.2.3 产业全图

从产品到行业再到产业，越发宏观离产品越远，但却离做产品的出发点越来越近。针对不同的产业，可以从一些财经资讯或报告中着重关注以下两点。

产业态势

我们常听说朝阳产业与夕阳产业的说法，这便是对产业态势的一种描述。实体经济与虚拟经济也是对产业的大分类。关注产业的走势，知道自己所处的产业处于一个什么样的状态，是在走上坡路还是在走下坡路，这些对于判断在特定时间采用激进抑或保守的产品策略很有帮助。

经济走势

产业经济是国家乃至国际宏观经济的一部分，了解我国以及海外市场所在国

的整体宏观经济趋势重在两点。一是大势，是上行还是下行，是衰退、复苏、过热还是膨胀，需要有一个基本的论调。二是政策，即所在市场可以享受的政策红利有哪些，要多关注多收集。

产品、行业、产业这三个全图构成了产品经理压箱底的石头，时常温故而知新，可以加深对为什么做产品的理解（WHY），也能够获得更为广泛的做产品的技术实现手段（HOW），并最终在产品功能上精益求精（WHAT）。梳理虽不易，砥砺前行之。

3.3 工作中如何学习

古代科学家也是哲学家。创立了解析几何的科学家笛卡尔说：“我思故我在。”率先提出原子论的古希腊科学家德谟克利特也说过：“莫让你的舌头抢先于你的思考。”由此可见思考的重要，工作中更是不例外，通过思考才能学习，这是不争的事实。

很多时候我们感到困惑，并非我们拥有的知识太少，也绝非思考不足。恰恰相反，我们面临的状况是知识太多，但知识之间产生的联系不够。

当微信中众多公众号的推送塞满一天的闲暇时光时，当订购的书籍占据书桌的大面积领地时，如果我们并没有感到更加充实，那很有可能是因为知识过载了。知识过载的焦虑，一方面来自于选择的焦虑，另一方面则来自于消化吸收的焦虑。

针对上述两种焦虑，一些伟大的智者已经给出了方法，让我们能够举重若轻地在工作中进行思考，这种方法就是“广泛阅读，充分联想”。

广泛阅读

根据联合国教科文组织若干年前公布的数据（<http://www.worldometers.info/books/>），全球每年出版的图书逾百万种，对于一个人来说，想要阅读完所有的书几乎不可能。即便我们面对的书籍集合缩小至母语类别，读万卷书对绝大多数人来说也是痴心妄想。面对这样的情况，选择什么样的书就变得尤为重要。有的人倾向于阅读更为实用的书籍，如技术书；另一些人则倾向于阅读虚构类的书籍，如小说；也有一些人会兼容并包，来者不拒。如果让我来选，我更倾向于赞

同所读书目种类较杂的人的观点。

诚然，许多书读了之后，并非立刻能用得上。但我仍坚信泛读与杂读对学习的帮助，因为这是产生联想的基础。如果时间困住了我们行万里路脚步，不妨让万卷书带我们打开想象的天窗。郭沫若先生曾经告诉过科学工作者，不要把幻想让诗人独占了。

充分联想

在广泛阅读的基础上，联想才会产生素材。特别是在跨领域联想方面，不得不推崇当代两位不同领域的代表人物，一位是投资界的查理·芒格，一位是文学历史界的威尔·杜兰特。

在《查理·芒格的智慧》一书中，作者并没有介绍资产负债表或者股票估值的任何公式，恰恰相反，作者回顾了物理学、生物学、社会学、心理学、哲学、文学以及数学等各个学科中的思维模式，并将其与投资领域的理论建立联系。看起来不相关、不搭调的若干个学科就这样被撮合、联系在了一起，本质上是基于“表面不相关的不同学科其基本真理是相通的”这一事实。承认了这一点就不会好奇物理学中的均衡与市场中的均衡为什么用的是同一个词汇，也不会对生物学中的物种进化与市场中的演变这两者间的相关性有任何怀疑。在他的理论中，这个思维模型被称为“格栅模型”。

杜兰特夫妇所著的书籍《历史的教训》曾是政府官方推荐的书目。在书中，我们可以看到与芒格如出一辙的广泛涉猎。从地质的历史讲到生物的历史，从种族的历史讲到性格的历史，除此之外还有道德、宗教、经济、社会、政府以及战争在历史中的作用。究竟需要何等的长期阅读、积累与萃取才可以写就这样一本薄薄的小册子啊！在其面前，许多大部头的书都黯然失色。

工作中的学习并非一蹴而就，也非一劳永逸。许多当下看来不会带来即时回报的学习终将在未来产生效果。如果把学习也看作投资，那么不仅要看看当下能否直接指导实际工作（短期回报），也要看在未来能否有更长远的价值（长期回报与现值）。

part two

第二部分

古往今来的数据思维

第 4 章

历史中的数据思维

- 4.1 人口普查：最早的数据埋点策略 46
- 4.2 命令与征服：可视化最早的用意 49
- 4.3 科技革命：助力数据产品落地 54
- 4.4 数据驱动决策的历史溯源 57
- 4.5 管理咨询：使用数据降本增效 61
- 4.6 聊聊统计学 64
- 4.7 LEHD：美国的第一个大数据项目 67
- 4.8 历史给我们数据思维的启示 69

恩格斯在《自然辩证法》一书中提到：“思维着的精神是地球上最美的花朵。”一方面指出了思维揭示事物本质和规律的作用，另一方面也展现了人们在意识的指导下能动地改造客观世界的神奇能力。

数据思维作为思维的一种，可以揭示事物规律；数据产品则是人们能动地改造客观世界的工具。本章中的历史故事受涂子沛先生的著作启发，经过加工、整理与充实而呈现。数据思维并非昙花一现，一朝一夕；也绝非大力丸立竿见影，但坚持秉有数据思维的确可以让我们渐进地发现改变。

正如涂子沛先生在其著作《数据之巅》中说的一句话：“拉长历史的镜头，我们可以看到，一件事情本身的成败并不是故事的全部，因为一事将牵出另一事，万事万物总是互相关联效力，只要事情代表了未来的发展方向，就一定会以某种方式结出果实。”

下面就让我们一起走进数据思维。

4.1 人口普查：最早的数据埋点策略

说到埋点，产品经理和技术人员一定不会陌生。数据埋点是数据收集的手段之一，也是从事数据产品工作或数据运营工作所绕不开的一个问题，毕竟“巧妇难为无米之炊”，要想进行数据分析必先有数据。

4.1.1 埋点的技术视角

技术人员谈论埋点一般包括埋点技术、埋点内容以及埋点位置等。作为一名在数据道路上刚刚出发的产品经理，对这些有大致的了解是必要的。许多技术博客与论坛中已有详细介绍，这里仅做简要概述。

埋点技术

埋点技术主要分为代码埋点、可视化埋点、无埋点三类。代码埋点是指在网页面中嵌入一段 JavaScript 代码或针对某个点击事件单独编写代码，从而监控访问网站的用户行为，如 Google Analytics、百度统计、友盟等都采用此类方案。可

可视化埋点是指，在被监控页面中嵌入代码后，用户在另外一个可以展示被监控页面状态的界面上，通过点击交互的形式，指定需要监控的事件与页面元素，进而获得用户行为。而无埋点则可以理解为“全”埋点，即通过代码嵌入的方式智能地搜集该页面全部元素与用户点击行为。这样看来，埋点都使用代码来完成，只不过方式各不相同，代码埋点应该是写代码埋点，可视化埋点则是可视化交互埋点，而无埋点更像是代码智能全监控。

埋点内容

无论采用上述哪一种埋点技术，埋点采集回来的都是一个一个的事件，简单来说就是某人、某时与某事组合而成的一次事件。通过这种细节数据的累计，就可以获得诸如某件事有多少人进行、某人在一周内做了多少件事等汇总统计任务。与业务结合起来，就可以知道人次或事件发生的次数、两件前后相互连接的事件之间的转化率（下单群体与下一步付款人群之间的比值关系，浏览网站人群与随之登录人群的用户比值等）、留存比例（经历了一周或一个月后，还有多少比例的用户留下来，被广为人知的是同期群分析 Cohort Analysis）、时长（用户在一个页面停留了多长时间）、内容（用户看了哪些内容）等。

埋点位置

有了技术也有了实际的内容，接下来需要考虑的就是在什么位置进行埋点。我们通常会把做技术的同学分为后端与前端，粗浅理解起来可以这么认为：前端的工作成果用户看得见摸得着，后端的工作成果用户往往用起来却不易察觉。于是做界面开发的同事离用户更近，故而被称为前端。埋点的位置也相应地可以分为前端与后端两个地方，上述埋点技术更多的是前端埋点技术，后端埋点技术可以认为是从数据库或服务器的日志中“扒”数据。两者各有优劣。前端可以把用户的细节行为尽揽囊中，且不需要联网进行收集，也正是由于这些特点，使得收上来的数据不能及时传送给产品分析人员，时有丢失，造成对业务分析的不及时或不准确。后端埋点尽管可以保证及时性与准确性，但是对于用户的细节行为却一筹莫展，更别提一些如小游戏等应用在非联网状态下的数据收集了。

4.1.2 埋点的时机与策略

埋点的技术与位置或许和技术人员相关，可是具体针对什么内容埋点，哪些埋哪些不埋，哪些先埋哪些后埋，埋好点的数据具体有什么用，却是产品经理需要考虑的事情。

针对这个问题，往往存在着两种观点：一种观点认为，不管三七二十一，虽然不知道数据有什么作用，先收回来再说；而另一种观点则认为，按需收集，小步快跑，快速迭代，需要多少收多少。应该说，两种观点各有道理。

不同的发展时期需要采用不同的数据收集策略。如果是初期，小步快跑在理论上是对的，毕竟这个时期的重要任务是保证 MVP（Minimum Viable Product，最小可行产品）的发布并且根据准实时数据动态调整产品的设计与规划。而在成熟时期，企业在市场中的布局已经较为完备，可能还涉及总公司与分公司、并购企业之间的协作问题，若要达到管理的目的，需要以详实的数据为抓手。

要分清这两个时期何其难也，它们本身就是孪生的一对。成熟时期中往往也会有新产品的规划与发布，微观来看似乎只需要按需收集即可，但是宏观来看又处于成熟时期，需要多多益善。因此，收集的策略究竟是严格还是宽松就很难定夺。另外，对于数据收集来说，必定是以结果为导向的，究竟可以发挥什么样的数据价值应该是收集数据之初考虑的。可是这本身就是一个“先有鸡还是先有蛋”的问题，也许可以事先定出优化产品设计的目的从而进行按需埋点，但是对于研发数据产品的同事来说，没有数据就很难做出规划，如果没有人向前迈一步，这将很大概率成为一个死锁问题。

当我们困惑的时候，不妨跳脱出来，将目光投向深邃的历史长河，向历史求答案。在涂子沛老师的《数据之巅》一书中有更为详尽的史实，这里仅做符合主题的摘要与转述。

这得从美国的建国历史开始说起。美国在立宪之初就决定了参众两院的国会制度，其中众议院议席的分配原则基于各州的人口比例，那么统计各州人口、了解州情数据就成为了重中之重。一方面，人口多可获得更多的众议院席位从而有更大的话语权，另一方面，人口多的各州也需要向国家缴纳更多的税。在这个权利和义务相互制衡的条件下，就需要有一个客观公允的第三方来进行此项工作。

我们可以想象的公正客观的第三方有两个：一个是国会自己管理的机构，在各州的监督下进行；另一个是独立于各州与国会的民间机构，这个机构中充满专业且没有利益纠葛的人士。美国在实施普查之初就选择了前者。

从 1787 年立宪会议开始，美国决定在每个 10 的整倍数年进行人口普查。因此第一次人口普查定在了 1790 年，这一年的人口普查仅仅以家庭为统计单位，填写家庭的人数、性别、种族和年龄。

在后来的 1820 年第四次人口普查中，又增加了职业这一项。了解职业分布有利于国家掌握劳动力分布，从而更有效地规划劳动生产政策。到了 1830 年，又增加了统计残疾人的选项，十年后的 1840 年，又开始统计文盲、精神疾病患者和牲畜的数量。到了 1850 年，对家庭进行的普查改成了对个人的普查。而随着普查的进行，人口的增多，入户的调查改成了邮寄调查，逐个调查改成了抽样调查，手工统计改成了机器统计。

历史是一本沉默的书，但当翻开它的时候，它却又滔滔不绝地给你伟大的启迪。从美国的人口普查中我们可以得出，任何一次数据收集内容和手段的变化，都是符合当时的具体情况的，没有脱离实际目标的手段存在。所谓“以终为始”，在美国的人口普查过程中体现得淋漓尽致。反观我们现在的数据收集，在收集之前或许得想明白，我们大致能用数据来做什么？确定了目标再进行收集，可以起到事半功倍的效果，同时也更好地有的放矢。

4.2 命令与征服：可视化最早的用意

还记得初中时看《名侦探柯南》动画片，剧中柯南利用阿笠博士为他发明的犯人追踪眼镜（Criminal Locator Glasses）进行追踪、定位、监控，功能强大，无所不能。原本在动画片中的场景，如今已经变为现实，微软的 HoloLens，谷歌的 Google Glasses，以及各大厂商分别发力研究的 VR（Virtual Reality，虚拟现实）、AR（Augmented Reality，增强现实）产品，让我们有了更为真实和沉浸式体验，也让我们对未来充满了期待。

4.2.1 可视化大家说

VR 与 AR 从广义上来说是可可视化的一种形式，然而可视化却包罗万象。让我们简要地从技术、类型以及领域三个方面来快速了解可视化。

技术

技术是为了解决特定领域的特定问题而产生的。科学工程领域、金融工商业领域、娱乐生活领域需要的可视化是完全不同的。科学工程领域包括了地理遥感、航空航天、医药生物等方向，其过程中不仅需要记录数据，还需要对几何外形、拓扑结构、纹理形状等进行分析，这涉及了多维度的数据，学术上有一种说法称为张量场，其中涉及了微积分、场论、拓扑数学等。金融工商业领域关注的可视化技术粗浅说来便是图与表，也是我们接触最为频繁的可视化，除此之外还有文本与地图等。这里需要的可视化技术包括了信息组织与呈现形式方面的技术，如时空可视化、层次与网状数据可视化等，学术上的概念实在不宜堆砌过多。最后的娱乐生活领域则需要人机交互与计算机图形学技术，我们使用的体感与 VR 等技术便属于此类。

类型

根据数据的类型我们也可以对可视化进行分类。数据大体说来主要分为一维、二维、三维与多维。一维的数据就是数据点，如某个时间点公司的股票数据，或者某次实验的测量值。这样的数据呈现方式往往是与时间耦合的，或者以比较的形式而存在，如股票的变动与时间进行了耦合，又如研究型论文中的实验数据比较则是以比较的形式存在。二维数据是指在二维平面图上进行展现的数据，例如商业领域要进行竞争分析或市场环境分析时，往往会采用波士顿咨询公司发明的 BCG 矩阵来进行相对市场占有率与销售增长率两个维度上的分析。三维数据是一种以 3D 方式展现的数据，如城市的热力图，卫星遥感得到的地理建模图形等。多维数据的展现则更多取决于展现形式的设计，很多巧妙的方式可以在一个二维象限中融合五维甚至更多维度的数据。

领域

尽管可视化是一门与图形学相关的技术，但它却不是图形图像学研究者的专利。谈到可视化的时候，大致有这样一些人会认为这个概念与自己的职业有关联：

首先是 IT 人员，特别是一些前端开发人员，他们需要使用工具进行数据展示，他们管可视化叫 Visualization，即将数据通过直观的方式展示，使用的工具是如 D3 或者 ECharts 之类的基于 JavaScript 进行编码的程序包。其次是设计人员，数据可视化对于他们有个可爱的称呼，叫作 Infographics，即信息图，也就是通过一张精美的图片展示数据的结果。通常我们在一些微信公众账号上看到的 Q 版数据展现形式都可以称为信息图。还有一类人员对于可视化比较有感觉，就是决策者，他们对于可视化的理解是报表，即 Dashboard。这里的可视化仅仅是报表的一种组件或者部分数据呈现形式，他们更关注的是从信息中进行判断获得下一步的行动准则。

4.2.2 可视化与历史

关于可视化工具的文章比比皆是，每篇文章都条理清晰，分类明确，那么可视化究竟是怎么产生的呢？

如果在搜索引擎中键入“可视化案例”，我们会得到诸多可视化的实例，有当下的，也有历史上的。其中诸多页面以“人类历史上最重要的 X 张可视化图片”为标题吸引你的注意力。而当键入“可视化起源”的时候，除了能找到《哈佛商业评论》英文原刊 2014 年 6 月（HBR.org）的一篇文章（<https://hbr.org/2014/06/the-story-of-the-first-charts-in-three-charts>），其余的似乎并不能很好地解释这个问题。而且 HBR 也仅仅给出了折线图、柱状图和饼状图的历史溯源而已。这使得我们更加好奇可视化是如何产生的。

最早的可视化可以追溯到公元前 600 年的古希腊，阿那克西曼德（Anaximander）是当时的天文学家和哲学家，那个时候的希腊居于爱琴海沿岸，其西北角是欧罗巴的大片地区，东北角是亚洲的大片地区，而东南边则是迦太基（利比亚当时属于迦太基），为了描述这样的政治地缘格局，阿那克西曼德绘制了一张世界地图，如图 4-1 所示。这应该算是最早的世界地图了吧。站在地图背后的，是帝国执政者统治的欲望，因为只有掌握了世界的模样，才知道自己下一块要征服的土地在哪里。

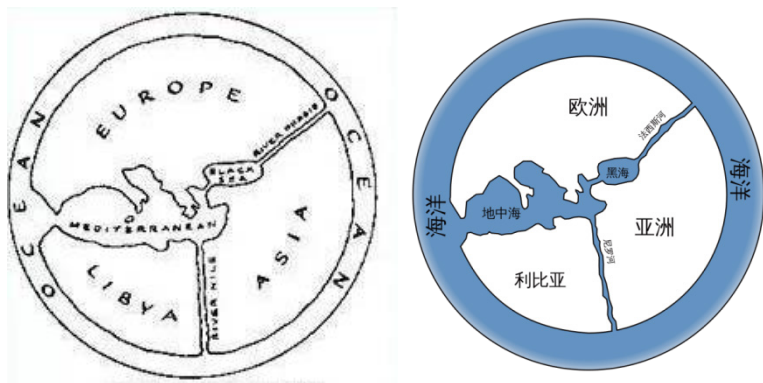


图 4-1 阿那克西曼德绘制的世界地图

扩展阅读：你看到过哪些有意思的地图？（知乎专栏）

战争有三个要素，分别是权、钱、人。权靠封赏，人靠抓丁，钱靠赋税。那么钱是如何征收的呢？显然是按照人头！因此了解各个地方有多少人口便成为关键。在美国南北战争期间，为了帮助了解各地的人口，美国的人口普查员使用了可视化的方式展示数据，涂子沛老师在书中展示的正是这个事例。

大致是在同时期，英法对俄宣战，经历了近代史上的第一场现代化战争克里米亚战争。这次战争中首次出现了蒸汽船、榴弹炮、来福枪、电报和天气预报等先进武器及技术，也正因为如此，战争死伤格外惨烈。大家可能对这场战争并不了解，但一定听说过一个护士的名字，南丁格尔。她除了参与克里米亚战争救死扶伤之外，还发明了一种可视化的图标，称为南丁格尔图，如图 4-2 所示（图片来源于 Wikipedia）。这种图使用极坐标的方式展现战争进行的不同时期各类伤亡情况占比。

也不全对。在欧洲启蒙运动时期以及近现代，可视化技术被用来进行艺术创作。正如我们在各种网站上看到的用可视化图形的形式展示拿破仑在 1812 年攻打俄国时从进军到败退的场景，如图 4-3 所示（图片来源于 Wikipedia）。虽然描述的是战争场景，但作者做出这张图已经是 57 年后了，算是使用了艺术的眼光来看待战争。

除此之外，Joseph Priestley 于 1765 年绘制了甘特图，我们现在用它来进行项目管理，但当时只是一种艺术创作，用来说明历史上名人的生活时期。Charles de Fourcroy 于 1782 年创作了面积图，显示欧洲城市的发展进程。苏格兰工程师和经

一抹趣味的色彩；另一方面，是社会需要促成了可视化这项有趣的技术发明。从古时帝王用于号令天下的地图，到近现代解决社会事务与征服科学难题的实用工具，不得不说，命令与征服，与可视化一路相伴。

4.3 科技革命：助力数据产品落地

数据产品是在大数据时代被写入各家企业战略的又一个新名词，一如当下火热的人工智能。谈概念，必先谈认知。知乎上有位回答者“老读悟”给出了他的理解：“数据产品是可以发挥数据价值去辅助用户做更优决策（甚至行动）的一种产品形式。”广义上来看，像谷歌、百度、360、必应等搜索引擎，今日头条等个性化新闻推荐引擎，豆瓣音乐等音乐推荐 App 都算是数据产品，只不过因为产品的产生早于大数据概念的火爆，所以一般不会被划分到数据产品中。而狭义上的数据产品则是指数型、报表等商业智能系统或客户管理系统等。

无论是其广义还是狭义的定义，我们都可以从中找到一个共通点，即消费数据。如果数据只是呈现，而没有人消费，则算不上真正意义上的数据产品。从这种对于数据产品的认知来说，人口普查就是一种数据产品。首先，没有人会否认人口普查的数据性质。每次普查过程中，需要拟定指标，找准范围，然后进行计数。其次，人口普查是为经济决策等社会管理工作服务，其消费者更多的是政府以及科研院所等机构，故而也满足消费的定义。所以，我们可以以人口普查作为数据产品的切入点，洞察技术革新与数据产品落地之间的关系。

现代人口普查距今已经 200 余年，学术界的共识认为现代意义上的人口普查始于 1790 年的美国，与此同时美国也将人口普查写进了宪法。现代的人口普查大致也分为三个发展阶段，从 1790 年至今，以 1870 年与 1950 年为两个时间分割点进行划分。1870 年与 1950 年发生了什么呢？为什么要以这两个年份为分割点划分成三个阶段呢？

19 世纪 50 年代，全球陷入一片战争，各国列强竞相争夺全球利益，各个资本主义国家相互之间明争暗斗。可以说从 1790 年到 1870 年这段时间内，各个资本主义较为发达的国家开始进行人口普查，主要目的是了解本国及殖民地区的劳动力状况与资源情况。

而对于 20 世纪 50 年代来说，二战刚刚结束，各国利益重新进行了分配，在这个过程中，战前统计主要是为了了解新兴市场的劳动力与领土情况，战后普查主要是为了了解人口资料状况而恢复国家的生产。

在这个过程中人口统计经历了手工统计到机械统计再到电子计算机统计三个阶段的数据落地方案升级。下面从这三个阶段分别介绍美国人口普查过程中数据落地的进展。

4.3.1 手工统计

从 1790 年开始美国进行了立宪之后的第一次全国规模的人口普查，其目的是解决众议院议席在各州之间的分配方案与各州交税义务等问题。随着时间的流逝，到了 1850 年和 1860 年的第 7 次和第 8 次人口普查，人口逐渐增多，统计也从以家庭为单位变更为面向个人，仅以这两次人口普查来说，点人头的时​​间花费了半年，而汇总分析数据则分别花费了 9 年和 6 年。且不说能否指导国家政策的制定，数据能否准确反映当下的问题都是个问题。

效率低下的根本原因是体制问题，当时负责人口统计的人口普查局隶属于国会，基层普查员的管辖权力在于联邦政府，因而造成了低效率。除此之外，人口普查人员的政府背景被民众所提防，使得统计数据并不准确。出于数据准确率与统计效率的考量，1869 年，加菲尔德提出普查队伍的专业化，从而以一个近似于第三方的角色来替代政府进行普查。这个建议在众议院通过，却遭到了参议院的反对。就在同年，为了统计员工工作期间权益，调和劳资矛盾，波士顿率先提出成立劳工统计局。1870 年，加菲尔德提名弗朗西斯·沃克主持普查工作。

4.3.2 机械统计

在沃克主持 1870 年人口普查的过程中，沃克的下属查尔斯·西顿发明了西顿制表器，利用该机器可以提高比较数据的速度。1879 年，加菲尔德当选为总统，于是加菲尔德如愿以偿地将普查的队伍进行了专业化，并推动了基层普查队伍受联邦辖制的权力。1880 年，美国迎来了第 10 次人口普查，沃克理所应当受到加菲尔德的信任主持此次人口普查。在这次人口普查中，除普查人口数量外，还普

查了出生率、死亡率、农业、社会以及工业等情况。

1881 年，沃克雇佣了霍尔瑞斯来到 MIT 任教，这个既懂得统计也懂得机械的跨界青年霍尔瑞斯明白，属于他的时代即将到来，而跨界的力量也即将显现。

1888 年，多么吉利的数字，霍尔瑞斯不负众望，发明了“打孔卡片制表机”，通过精巧的电路设计操纵机械来对数以万计的卡片上的数据进行统计。你或许会有两个疑问：那个时候就有精巧的电路了吗？既然使用了电路，是否就属于电子统计了呢？

第一个问题很好回答，爱迪生发明电灯的时间是 1879 年，因而 1888 年应用电路并不稀奇。对于第二个问题，我们界定电子统计和机械统计的原则是，数据究竟存储在物理介质卡片上还是电子磁性装置中。虽然使用了电路，但是依然属于机械统计。图 4-4 是位于硅谷的计算机历史博物馆中展出的老式打孔机进行卡片打孔操作的情景，很好地显示了当时记录数据的方式，以及进行数据录入和采集的过程。跳出历史的沟壑，再俯视滚滚历史的洪流，不得不佩服为这个时代带来新科技和新发明的杰出人士。

1890 年，美国便开始在普查中应用该机器。不禁要佩服霍尔瑞斯的眼光，他在发明当初就开办公司创立企业，从而在接下来的时间中将该项发明应用于统计而大赚了一笔。

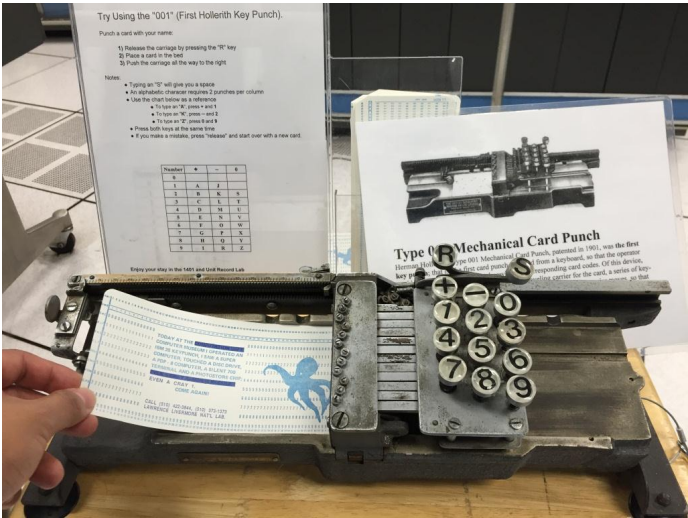


图 4-4 硅谷计算机历史博物馆中的机械打孔器

1909年，霍尔瑞斯的专利到期，他公司的红利期已经过去，各家开始出现竞争。由于经营不善，霍尔瑞斯的公司被迫和其他公司合并，成为CTR公司。而担任CTR首席执行官的正是沃森。CTR就是IBM的前身，正式更名是在1924年。一家企业在红利期内可以大肆通过toG的项目来攫取利润，然而通过技术垄断获得的壁垒还需要和内部的管理进行博弈，当管理的代价抵消红利的利润，就是企业的转折点。这值得被所有的企业铭记，不能重技术而轻管理。

4.3.3 电子统计

时间来到1946年，世界上第一台通用计算机ENIAC诞生了，起初这个计算机只用于军方项目，而两位参与该项目的科学家却没有获得好处，于是出走创办公司。在接下来的1951年普查统计中，美国就使用了这两位科学家发明的UNIVAC计算机进行数据处理。

IBM当时是落后的，他们只能使用制表机这种老式的机器，自然失去了许多客户。于是，IBM下定决心转型为电子计算机制造公司，于1953年制造出了701机器，并被市场接受。从此拉开了近现代计算机时代的帷幕。

数据的落地经历了从人工统计到机械统计再到电子统计的三个阶段，宛如一个新生儿从孩提到青年再到中年，不仅有成熟的思想与技术，还为这个社会的进步奠定了基石。这意味着我们作为后人可以站在巨人的肩膀上，向更高的目标进发。而进发的过程中我们要做的必不可少的事情便是以史为镜。

4.4 数据驱动决策的历史溯源

当我们在高喊“数据思维”的时候，往往身体还滞留在凭借感性和经验做决策的阶段。有的时候我们会莫名地升起狐疑，为什么明明已经迈开了步子，但是却还没有发生位移？原因只有一个，“知行合一”做得不够。

从生物学上讲，当我们高喊口号的时候，大脑皮层进行着自顶向下的决策，指挥身体的各个部分进行协同工作；而与此同时，经过几百万年历史进化而来的杏仁核却抵触着这种决策，它们会选择最为安全的方式让身体协同，并自下向上

地将身体的感受反馈给决策者大脑皮层以影响其决策。对于利用数据思维这件事来说，我们生活中固化的经验就是安全的，因为“因循守旧”至少不会触怒权威，可以与周遭和平共处，最重要的，对于大脑来说，不使用数据思维进行思辨，就不需要消耗更多的 ATP 能量，因而这是一种安全的方式。

而利用“数据思维”则是一种冒险而激进的方式。首先是能量的消耗增加了，如果用人脑类比计算机，那么杏仁核就是烧制了固化程序的芯片，由于这些经验是被固化的，自然其开发成本以及各项性能都已经在体系中得以优化，如果这个时候要让自己利用数据思维来进行工作，那么无异于为计算机增加了一个定制化的芯片，这样的代价以及各种能量的消耗不言而喻，这对于计算机与人体系统都是一个不小的冒进，这也可以算是人们常说的“走出舒适区”的一种。其次是使用数据思维需要探索未知的世界，没有更多的先例进行指导，我们只能参考少得可怜的行业案例照猫画虎。然而当遇到实际情况的时候，往往千差万别，这个时候运用数据思维不可能阅万卷书以小心求证，这无异于一万场小型的创新，非常需要勇气。

喊得响并不一定能做得漂亮，究其本质是惰性。因而要建立数据思维，一要身体上不怕累，二要思维上保持勤奋。

当我们谈到大数据或数据思维时，必须以特别虔诚的态度将目光投向西方，特别是太平洋彼岸的美利坚合众国（之所以不说是美国，原因后面再解释）。很多科技发明与创新源自它西海岸一个从旧金山到圣何塞的狭长湾区——硅谷；而经济财富的运转与交易则发生在它东海岸一个与自由女神像仅仅相隔一河的短促大街——华尔街。下面就让我们追溯美国的历史来看看美国的仁人志士是如何用数据来驱动决策的。

4.4.1 美国建立时用数据分权

1776 年，美洲大陆的人民摆脱英国殖民的侵害，申诉自己的权益，于 7 月 4 日宣布独立，但直到 1783 年才被承认。他们原本的信念就是自由，因而在独立的过程中是无政府的，大家“各自安好”，“安居乐业”。然而好景不长，大家发现这样的生活状态并不开心，更不富裕，于是开始平心静气地坐下来讨论立宪的问题，建

立一个强大的政府来管理国家。这个时候，时间的年轮已经滚到了 1787 年。

在立宪的时候，有一个很具体的问题，当时在美洲大路上大家各自占山为王，分为各州，这些州的代表们聚在一起开会讨论立宪问题，同时解决一个很棘手的问题，即“各州谁说了算”。在立宪的过程中进行了三大妥协，这里只说两点。

首先是被称为“伟大的妥协”的参众两院制度。提出这个问题主要是因为各州主张不一致，用参众两院制度可以平衡，众议院中人数多的州占得席位多，而参议院中每个州都是两个席位，大家平等。任何决议一定要参众两院批准才行，参众两院形成了国会制度。

其次是“五分之三妥协”。赶走了英国殖民者，那么每个州要计算人数时，黑奴该怎么算呢？如果一个黑奴算一个人，那么白人肯定不同意，这等于剥夺了他们的尊严。最后达成一个共识：使用一种折中计算方式，一个黑奴等于五分之三个白人，这就方便计算各州的人数了。

看似已经达成一致，其实各州暗流涌动。主要原因在于，众议院既然是按照各州人口比例进行分配的且众议院席位有限，那么就肯定会出现每个州分到的席位存在小数的情况。围绕这个问题，领导联邦党的汉密尔顿（汉派）与领导民主共和党的杰弗逊（杰派）都给出了各自的数据计算方式，之后亚当斯、韦伯斯特以及哈佛学派的亨廷顿都围绕杰派的计算方式给出了自己的修订版本，其目的都是解决由于人口动态增长使各州席位分配不均而产生的矛盾问题。具体的内容就不介绍了，感兴趣的可以阅读《数据之巅》。

上述故事可以见得，美国从建国之初，就理性地使用数据来进行讨论，而不是充斥自身与所代表州情感的感性论战。这可以说为美国后来的数据驱动决策思维开了一个好头。

4.4.2 南北战争时用数据进军

时间来到了 19 世纪 60 年代，美国发生了南北战争。一言以蔽之，南北战争就是北方希望发展资本主义经济取代南方的种植园经济，从而围绕南方究竟该不该废除黑奴而打的一场战争。

总统林肯通过美国 1860 年人口普查，得出 69% 的青壮年归于北方，估计北方

会赢得战争。而事实上，尽管北方有优势，却是南方于 1861 年 4 月 12 日先挑起战争，原因就在于南方掌控经济。

林肯为了摸清楚全国的经济作物情况（那个时候美国还是个农业国），于是在次年 5 月成立了农业统计局，掌握全国的农业态势。在战争过程中还涌现出一个英雄般的人物——谢尔曼将军。谢尔曼将军是北方军团中东线作战的将领，他于 1842 年 9 月占领了南方重要城市亚特兰大并实行了“三光政策”，摧毁了南方的经济。11 月，他依据当时的人口和农业等数据做出智慧的判断，决定丢弃辎重，轻装上阵，小步快跑向东挺进，这次向东挺进的战争就是著名的向大海进军（march to the sea）。次年（1965 年）4 月 9 日，南方重要将领罗伯特·李请降于北方将领格兰特，于是南北战争结束。

从战争中学习是士兵的本能，我们虽不是士兵，但却从林肯在开战之初通过数据判断战争形势，以及谢尔曼将军通过数据决定战争方略的智慧谋略与英勇胆识上看出数据的重要性。在战争前，以人口统计数据为依据判断战争走势；在战争中，通过经济统计数据作为指南选择战略要地。最终的胜利或许有歪打正着的成分，或许有其他因素的影响，但这种依靠数据做决策的行为却渐入人心，历久弥新。

4.4.3 经济发展时用数据裁判

时间继续挺近到 20 世纪初，已经没有了建国时期各州的纷争，也没有了南北战争时期的硝烟，这个时候美国经过 19 世纪末期 30 年的发展，已经在资本主义的道路上小步快跑，并且于 19 世纪末，准确来说是 1894 年，成为了世界上第一大经济体。

市场经济的问题无非就是商品市场、劳动力市场与资本市场的交叠问题，在三个市场自由运行的过程中难免出现各种各样的社会问题。譬如在劳动力市场上，劳工的处境岌岌可危，美国在 1907 年提出了“匹兹堡调查”，用一系列数据说明了劳工的窘境，成功吸引了人们的关注并促成了工伤赔偿制度的出台。再例如，1908 年，布兰代斯为超长时间工作的女工打官司从而促成政府规定每日的工作上限时长，当年的规定是 10 小时每天，而今天待遇越来越好，已经变成了 8 小时。

要知道，这一切正是布兰代斯使用法律分析与援引数据的双重武器换来的，也正因为如此，布兰代斯被称为“人民的变化律师”，他还提出了“阳光是最好的消毒剂”的监督政府策略。

除此之外，在美国的大建设时期，当工程兵团与农垦局为了谁来建设争执不下的时候，他们除坐而论道之外，还使用了在当今看来也非常时髦和前卫的“成本收益分析法”来进行决策。这也是1950年成本收益分析法“绿皮书”标准出台的最直接原因。可见量化分析是一切的源泉，这也是管理学大师戴明和德鲁克尽管有着剧烈的分歧，但在“无法量化就无法管理”这一理念上保持共识的原因。

数据驱动的思维，发端在我看来不仅始自18世纪的美洲大陆，如果进一步考证，会发现它可能植根于美国的文化母体——欧洲。那个时期的欧洲，经过了古典、中世纪、文艺复兴、宗教改革、启蒙运动已经来到了浪漫主义运动时期，抽丝剥茧，层层追问，会发现这个思维的根已经深深地植入欧罗巴整个社会中。我相信，如果你也愿意相信，从非洲大草原出走的第一批智者便已经孕育了这样的文化。既然如此，我们有何理由拒绝如此先进、理性的思维？又如何能与历史的进程为敌错失良好的发展时期？我们能做的便是学习与接收数据驱动决策的思维！

4.5 管理咨询：使用数据降本增效

想必大家阅读过无数有关大数据的文章，你一定看过很多产业政策文章（例如“某地十三五期间大数据产业规划案”），也一定阅读过各种信息化图表的文章（例如“某某马拉松大数据”）。除此之外还有大数据职业规划文章（例如“数据科学家的几个阶段”），又如一些大数据行业案例（例如“某某企业大数据实践”），当然还有大数据鸡汤（例如“关于大数据的这些，听听某某怎么说”）。

有人问我大数据究竟有什么作用，三个词语六个字足矣概括，分别是降本、增效、创新。这并非我的原创，很多经济学大师就“创新”这个问题发表了自己的见解，如熊彼特提出“创造性破坏”，又如克里斯坦森提倡“破坏性创新”。创新便是创造数据产品，除技术之外，还需要创意、灵感加上一些运气，并非时间用足就能够憋出大招。相较而言，降本、增效则相对来说有先例可寻，可以按图索骥。

不知道你平常是否有阅读报告的习惯，是 36 氪、虎嗅、艾瑞、TalkingData 一样的行业报告，还是类似于腾讯与阿里研究院以及各类国家智库发布的产业分析报告？抑或你已经开始阅读国外咨询公司的报告了？实话实说，越是向咨询公司靠拢，便越能够发现自己具备了把握当下命门以及长远展望视野的双重能力。而这两项能力是在进行企业管理的过程中必不可少的。

也许对大多数数据产品经理来说，上述段落纯属散侃，但如果我说数据产品的作用与企业的降本增效目的如出一辙，你应该不会反对。如此说来，组织结构调整的重新洗牌也好，市场战略梳理后的重新定位也罢，数据产品都是一种具体的实现形式。那么我们就不得不去了解管理咨询的那些事了。

4.5.1 咨询指引数据产品方向

现在每次美国大选之前都会有人放出风来，说自己预测总统大选的获胜者是谁。譬如特朗普和希拉里的竞争，很多公司的预测都有些偏差。不过敢于在答案揭晓前预测的都是勇敢的，还有一些公司马后炮似的拿预测准确做公共关系（PR）文章。当下的预测公司是一些以大数据和人工智能技术为核心的互联网或 IT 企业，而在近 90 年前，也就是 1930 年，也有一家公司尝试预测总统大选，便是盖洛普咨询公司。它的建立比我们熟知的麦肯锡咨询公司只晚了 4 年。

现在的数据收集已经是无孔不入了，但是在当时只能通过问卷来进行。盖洛普的首次亮相当属 1936 年成功预测罗斯福将战胜兰登赢得选举，这一爆炸性新闻使得当时看好兰登的主流媒体为之咋舌，从而也使得盖洛普一炮而红。在之后的过程中，盖洛普还成功预测了英国大选。总体来说，盖洛普的战绩还算不错，失利次数并不多，不过最近几次盖洛普表现不佳，从 2012 年大选后盖洛普已经不再进行民意调查了。

盖洛普除了使用民调数据来预测总统选举外，还使用该数据指导了电影《乱世佳人》的拍摄。当制作方为了“要不要拍摄这部电影”“选择哪个女演员担任主演”“采用黑白还是彩色来拍摄”等问题困惑时，盖洛普用专业的手段给出建议，并通过实践获得了认可。

无论预测总统大选还是指导电影拍摄，都是增加效率的一种体现。很多时候

的犹豫不决与裹足不前是导致决策失误与响应滞后的罪魁祸首，通过咨询公司的数据（当年是民意问卷调查，现在是大数据舆情）来进行反馈，或者利用行业情报进行分析，以做出最准确的判断与最及时的响应，这何尝不是一种先进的理念，甚至精益的思维呢？

4.5.2 管理启迪思维模式更新

什么是管理？每个人会因为自己的职位和处境不同而有不同的看法。对于涉世颇深的资深经理人来说，总能说出一番颇有哲学意味的话语。而对于数据产品经理来说，管理就是 KPI。说到 KPI，就要说到两个近现代管理学大师，一个是德鲁克，另一个是戴明。

每个领域都会有不同的学派，每个学派又有各自的追随者。德鲁克的理论强调以结果为导向的管理，也就是绩效管理，其理念风靡欧美，中国借鉴的就是德鲁克的这一套理论，这也是为什么走到哪里都会遇到 KPI 的原因。戴明则反对德鲁克的观点，戴明强调过程管理，也就是质量或流程管控，他的理念在欧美遇到阻碍，但却在日本得以发扬光大。欧美和日本都取得了辉煌，所以到底是德鲁克更厉害，还是戴明更实用，我们犹未可知，见仁见智。

既然当代管理上的很多方法与工具都源于日本，那么我们就来探究一下这个给日本企业洗脑的戴明博士究竟做了些什么。

1941 年，日本发动珍珠港事件，美国加入二战，为了国防工程的制造质量，戴明受委托对国防部门的 2000 名质量控制工程师进行培训。这些工程师在二战结束后成立了美国质量控制协会（ASQA）。

战后的日本谦虚至极向戴明求教，在 1947 年到 1950 年期间，戴明多次前往日本进行质量培训。在这个过程中，日本国力急剧膨胀，终于于 1960 年超过联邦德国成为世界第二大经济体。20 世纪 60 年代，日本以戴明的思想为基础，发明了质量管理七大工具，分别是核查表、数据分层法、控制图、鱼骨图、柏拉图、直方图、散布图。20 世纪 70 年代，日本又加以改进，发明了质量管理七大工具，分别是关联图、系统图、矩阵图、亲和图、PDPC 法、箭条图、矩阵数据解析法。

这一个个熟悉的名词紧扣我们的心弦，原来我们在规划数据产品、分析竞品，

甚至撰写报告时用到的方法起源于海对面的日本，而这一切又源自于更加遥远的太平洋彼岸的一位美国老先生。我们是否该好好思考一下德鲁克和戴明的理论到底哪个更加适合我们的管理呢？或者两者都存在，只是在不同的阶段使用而已？

管理是一种将人集中在一起工作和创造的技巧，理所应当以降低成本为首要任务。更为宝贵的是日本依据戴明的管理理论研发出的方法与工具共计 14 种，这 14 种工具与方法已经超越了管理本身，成为了数据产品分析与规划过程中的组合拳，所以其难能可贵之处就是对我们已经僵化的思维模式进行了一次升级改造。

既然管理咨询带给我们的不仅仅有数据产品落地的方向性指引，还有在梳理业务过程中思维模式方法论层面的指导。那么你还犹豫什么？打开浏览器，进入几个咨询公司的网站，下载几篇报告找找数据产品的灵感吧。

4.6 聊聊统计学

如果我问你美国的英文单词是什么，你可能会回答 America。那么全称呢？是 United States of America。这与统计学有什么关系呢？

关注上面的单词 state，它表达的意思是国家，而统计学的英文则是 statistics。有没有发现很相似的地方？是的，它们有相同的词根：stat。而之所以有这样的相似，原因就在于统计学的起源与国家有关。

我们无时无刻不在沿着涂子沛先生的足迹谈论人口普查与统计：美国在 1790 年的人口统计中进行了数据分权，手工统计到机械统计再到电子统计的三次统计技术革新，都是为了国家进行人口统计服务。不要惊讶，这恰恰就是统计学的起源。从历史跨度上来看，统计学的发展大致经历了三个阶段。

4.6.1 政治算术

通常意义上认为统计学起源于 16 世纪的欧洲，说到算术政治，就不得不说一个英国人——威廉·配第（William Petty，1623—1687 年）。了解其出生的大背景有助于我们理解其动机以及科学的发端。配第出生在 17 世纪，17 世纪的欧洲经历了 15 世纪的文艺复兴运动和 16 世纪的宗教改革，已经开始了新阶段的启蒙运

动。在这个时期，欧洲各个国家采用科学作为第一生产力，乐于接受和采纳新的科学思想来服务社会与国家政治。

统计学在这个时候有了用武之地，配第提出用数字指标体系来描述国家的经济民生状况，并提出了国民收入的概念，这便是统计学第一次登上历史的舞台。由于其最初的作用是为政治服务，且基本原理就是计数，因而被称作政治算术。不要小看这样的计数，这在当时是先进的科学思想，我们之前说过的“数据驱动决策”便是由此萌芽的。

4.6.2 频率学派

与配第同时代的还有另外一个英国统计学家，叫作约翰·格朗拖(John Graunt)，他比配第早出生 3 年，却比配第早离开这个世界 13 年。现在学术界对于谁是统计学之父还悬而未决。

格朗拖在其 42 岁那年，也就是 1662 年，发表了专著，其中表达的思想可谓数据处理领域的开山创举。首先，他提出使用简约的表格方式来展示复杂数据之中的信息，这有些类似于配第使用数据指标来表征数据的思想；其次，他提出数据在统计时有异常值，应当予以警惕；最后，他还提出了数据频率稳定性的概念，有点类似于方差的雏型。很难想象，这些了不起的成就都是在那个时候产生的。

后人在其基础上研究了更多的数据性质，譬如高斯分布（正态分布）等。这个阶段都可以称为描述统计学阶段，即使用数据指标来描述数据的性质和内在信息。其之所以与配第的主张分开，主要原因在于其不仅仅在政治中使用，还在农业、天文学、娱乐、宗教布道中广泛使用。

真正使频率学派的影响力扩大的是一个叫作弗朗西斯·戈登(Francis Galton)的人，他是达尔文的表弟(另一种说法是表哥，总之是 brother)，是一个生物学家，是他开创了回归方法，将本来的数据描述性指标升级为可推断和预测数据的性质以及未来发展趋势。因而从他开始，描述统计学走入了推断统计学的阶段(statistical inference)。围绕这个方法进行扩展研究的组织称为频率学派，即通过观察数据发生的频率来推测产生数据的事物本身的性质。

4.6.3 概率学派

所持观点与频率学派完全不同的人是托马斯·贝叶斯 (Thomas Bayes)，贝叶斯出生于 18 世纪初，他提出了一个很有意思的观点。让我来打个比方，按照辩证唯物主义的观点，人的好恶是可以三七开的，那么频率学派认为，假如一个人做的 10 件事情中，有 3 件坏事，7 件好事，这个人就是三七开的人。而贝叶斯却反过来看，他认为这个人是个完整的人，不存在好与坏的概率，而是他做的每件事在被认定为好事或坏事时存在概率。我们预期是好事的可能性是 $7/10$ ，是坏事的可能性是 $3/10$ ，随着我们观察到这个人做的事情越来越多，这个可能性会不断被修正，这样的可能性就是概率。

简而言之，频率学派用于客观描述事物或人的性质，随着不断观察使这个描述越来越准确；而概率学派却是通过观察不停修正自己的主观感受，使人们对于客观事物的感受越来越准确。

这个现在看似已经稀松平常的观点，在当时却是一大创举。后世围绕贝叶斯开创的观点进行研究，形成了概率学派（贝叶斯学派）。

就这样，欧洲延续着古希腊严谨的科学范式向前推进，由于美国是欧洲的殖民地，所以自然而然地，这样严谨的科学思想与严格的数学推理在北美大陆也生根发芽。

等到这一思想传入东方的时候，已经是 19 世纪下半叶了。首先向西方学习的是日本。19 世纪 60 年代到 90 年代，日本进行了明治维新，提出脱离东方加入西方阵营，其本质上是放弃儒家思想而拥抱西方科技。

1872 年，明治维新时期，日本进行了第一次人口普查。人口普查似乎是一个国家使用数据驱动决策的最好反映。1882 年，日本出版了统计年鉴。而统计学被正式引入中国则是在 1902 年，一个叫孟森的中国人翻译了日本人横山雅男的著作《统计讲义录》，正式为统计学登陆中国打开了局面。

统计学发端于欧洲，传播于大西洋彼岸的北美，后又经日本传播至中国，无数的历史事实仿佛在说明一个道理：要么被动接受，要么主动寻求突破，谁拥有更高的格局、更广的视野与更开阔的胸襟，谁就能在竞争中获胜。

4.7 LEHD：美国的第一个大数据项目

大型项目的开展总是从政府开始的，这是因为大型项目需要集约的资源，而政府正是握有资源的一侧；另一方面是因为项目总是由需求拉动，政府相较于个人或企业来说，对于未来的关注更多，因而对未来的需求总是旺盛。第一个大数据项目正是这样一个由政府主导的项目。

4.7.1 信息逐步开放

信息是数据的高阶形式，在介绍第一个大数据项目之前，我们不妨向历史追溯，看看美国政府是如何一步一步围绕数据与信息进行开放的。

1945年，美联社主编库伯首次阐述了“知情权”这个概念，1953年美国国会草拟了《信息自由法案》。但是这个法案真正通过已经是1966年。这个法案规定了政府需要公开静态文件与材料。

老百姓们不满足于静态文件与材料的公开共享，于是1976年国会又通过了《阳光政府法》，公开动态决策过程。随着信息科技革命的更替与互联网时代的到来，1996年，美国国会通过了《电子信息自由法案》，公开经由电子计算机和互联网存储的信息。2007年，国会又通过了《开放政府法》，公开政府委托第三方机构的信息。这一波又一波的民众促进政府进行数据开放的潮流使得信息在不同阶层间流动，进而使得民智开化，社会透明，政府运行效率提高。

另一方面，在这个过程中，1962年卡森出版的《寂静的春天》引起了民众对于环保的关注。民众对于数据的关注由以政府为对象扩大到了以企业为对象。1986年，国会通过了《紧急计划和公众知情权法》，定期公布某行业的企业数据，利用数据来制衡企业。

信息的逐步公开使得民众对于数据与信息越来越渴望。在大数据时代中，政府逐步开放各行各业的数据已经成为常态。从纽约政府开放城市数据，到中国上海举办SODA比赛，开放的政府拥抱变化，接纳众多智慧的脑袋，变得越发创新。

4.7.2 大数据项目开展

还记得我们曾经说过的统计人口吗？是的，统计学最初就是用在人口计数上的。这点毋庸置疑，但是这样的人口统计是离线的，一方面是由于技术手段的限制，毕竟最开始只能使用报表和打孔纸带进行数据存储与计算，另一方面也确实没有必要进行在线计算，因为人口统计的目的是进行宏观政策的制定，而宏观政策往往相对稳定，所以对于数据些许的偏差容忍度较高。然而这一切都在 2001 年 9 月 11 日被打破，相信谁都会记得这个日子。恐怖袭击发生在美国纽约，人们产生疑问，到底 911 恐怖袭击中有多少人丧生？民众再一次对数据的公开提出了要求！

人们对这个问题问得越深，政府就愈发感到窘迫。仅仅根据 2000 年的人口统计数据无法进行估算，因为纽约是一个大都市，很多人过着往返于纽约与住地郊区之间的生活，而这些人口却记录在郊区的人口统计档案中。因此，仅仅以世贸大厦附近街区的人口统计进行估算误差必定较大。那么怎么才能知道有多少人呢？

似乎办法也是显而易见的，通过查询世贸大厦中公司的数据，并且把这些公司的社保数据拿出来，看看每家公司究竟有多少员工就可以估算了，再结合当地的固定居民数，两者相互融合就可以得出结论。可是问题又出现了，人口统计数据在联邦政府，而社保数据在地方，如何才能打通两者的数据呢？为此美国政府成立了 LEHD（Longitudinal Employment and Household Dynamics）项目。该项目始自 2007 年，用于动态人口统计（白日人口）。这个项目打通了原本井水不犯河水的若干个部门的数据，因此可以算是名副其实的大数据项目。

我们在实施大数据项目时常常遇到要将多部门多领域的数据打通的场景，早在当年，美国政府就遇到过。通过他们的实践经验，我们似乎可以得到以下结论：数据的贯穿与打通不在于行政命令是否通达，而在于打通数据的目的是否明确，需求是否急切。

我们讨论政府数据的时候一直会说数据孤岛问题，而美国政府这次将联邦政府数据与地方州数据打通，正是完成了这样的使命！关于大数据对政府治理的应用案例，后面会专门介绍。

4.8 历史给我们数据思维的启示

至此，我们已经了解了与数据相关的历史故事。大家从中获得了什么呢？总结一下，大致有以下四点。

4.8.1 用数据说话

从 1790 年美国使用人口统计数据进行众议院的分权，到南北战争时期谢尔曼将军使用数据指导战争直捣黄龙获得胜利，再到政府使用成本收益分析法来指导建设的决策，无一不说明了使用数据的必要性。

用数据说话，是从经验走向精准，从模糊走向量化，从感性走向理性，从粗暴走向文明。在数据产品经理日常的工作中，我们无时无刻不在进行着决策。故而，数据产品经理要有先斩而后奏的勇气，组织上也要有让听得见炮声的人指挥战斗的胸怀。那么如果授权给你决策，你是否能够使用数据来指导判断呢？

在最开始，不妨机械地模仿，死板地较真，随着时间的打磨与实践的增多，相信数据驱动决策的思想一定可以在你的心中落地生根。

4.8.2 向贤者取经

有句话说“设计做不好，全因看得少”。其实哪里是设计，每行每业都需要打开自己的视野，多看多想。数据产品经理如此，中高层管理更是如此。要不然何必兴师动众到处考察借鉴呢？

那么究竟看什么？当然是看比自己优秀的人，比本家公司产品更前卫的产品，比当前商业模式更靠谱的模式。西方作为孕育创新的摇篮，是每一个热爱学习的人绕不开的取经之地。从成本收益分析法到日本借鉴戴明的管理七工具与七方法，每个都是精巧的设计与精炼的总结。这样的方法论提供给你的是一套系统的思考方法。当然也允许提出自己的主张，但如果还没有屡试不爽，不妨多学习些别人的方法吧。

方法论如此，产品、商业模式、设计、技术皆如此。

4.8.3 渐进性创新

美国人口普查项目并非一开始就收集到所有的信息；“统计人才专业化”也是经历了数十载的变迁才最终尘埃落定。历史事实无不在说明，凡事不能一蹴而就。

在企业中，数据产品经理肩负着组织大数据战略落地的重任，自然是压力的末梢受众。《孙子兵法》中讲“谋定而后动”，然而在压力与急切的心情下，往往会做出“先动后谋”“边谋边动”的行为，这些行为无异于葬送产品、拉低效率、增高成本。要知道一个战略的落地既不能一蹴而就，也不能一劳永逸，管理者需要有战略上的耐心。社会的改革尚且如此，企业的战略转型又岂会有异？

4.8.4 需求创造供给

“需求创造供给”是经济学大师凯恩斯的主张，旨在说明，任何事物的生产都是由社会需求引起的。围绕这个主张，我们可以找到很多的例子。从数据可视化来看，战争对效率的要求促使人们将大量用于决策的信息展示出来；进行人口统计的技术随着统计规模的扩大逐步经历了从采用霍尔瑞斯制表机到 IBM 电子计算机的历程。

数据产品经理们对于产品的制造也当围绕着需求进行。可是什么才是需求呢？走访用户收集到的东西是需求吗？那只能是需要。

我想你一定很纠结到底什么是需求什么是需要，让我们换个名词来解释。用户的需要就是他说出来的话，而需求则是说这话的动机。因而我们说，收集回来的是需要，分析动机得到的才是需求。进行产品规划的时候，很多情况下是为了真正满足用户的动机。难怪产品经理需要研究一些心理学，这也是进行用户研究的重要性吧。切记，出去访谈回来列举若干项，这不是需求，是需要！你需要深挖，再深挖！

要让需求创造供给，而不是先进行供给，再去寻找需求。

第 5 章

行业拥抱数据思维

- 5.1 大数据从何而来 72
- 5.2 大数据的全球格局与中国面貌 76
- 5.3 大数据+ “治理与交通” 81
- 5.4 大数据+ “零售与金融” 84
- 5.5 大数据+ “体育与教育” 89
- 5.6 大数据+ “医疗与旅游” 93
- 5.7 大数据+ “农业与制造” 96
- 5.8 大数据行业成熟了吗..... 97
- 5.9 大数据在产业中的位置 103

“经济因人们对于未来的憧憬而繁荣”，这句话并不难理解。如果不是因为看涨一个公司未来的发展，你一定不会购入这家公司的股票；如果不是因为对市场前景的悲观，你也不会抛售股票转而购入债券。因此，是社会中的每个成员对未来的良好预期与乐观估计，使得人们愿意投入通货，而这些通货转而又被银行放贷出去，服务大众的消费，服务产业的建设。在一定坏账率的情况下，当大部分产出高于投入的预期时，经济就在一级一级的杠杆中不断被抬升。

人们对于未来的预期需要一个概念作为载体，大数据无疑是让人们看到未来的新载体。新的概念最初萌芽于学术界，而后兴盛于各行各业。本章中我们将首先介绍大数据的发展历史，然后从地域的维度，以及行业的维度诠释行业与大数据的结合，最后将大数据摆到整个产业中，了解我们所处的位置。

5.1 大数据从何而来

IT 行业“造势”的能力比较强，从互联网革命到云计算，从大数据到人工智能，每一次新技术的兴起，总能用三五个字概括精炼，然后在“坊间”传播。当堵在路上的你打开收音机，会听到主播用大数据进行路况播报；当 CCTV 介绍节假日返程客流时，主持人也会说全国客运统计是大数据技术。似乎大数据已经和统计化为等号，但学术界也许对这个认知颇有微词。

很多人会问大数据和人工智能之间有什么关系，其困惑之处在于，对于同样一个任务目标，有的媒体鼓吹大数据技术的作用，有的专家诠释人工智能科技的魅力。大众对于这种二元的认知往往会出现失调，因为大家并不能够接受一个结果由两个原因所导致。其实大数据中也有人工智能，人工智能中也有大数据，这便让给出解释的人苦恼起来。

如果一定要对大数据和人工智能做出区分，我倒是倾向于将互联网、云计算等 IT 技术融为一体来说道说道。我们来打个比方，假设有一个制造企业的工厂，为大众提供消费产品或服务。为了建造这个工厂，首先得有土地、土建、通水、通电等，这些属于基础设施，互联网就是 IT 行业中的基础设施，几乎所有通过计算机呈现的 IT 服务都离不开那一根网线。为了生产产品，工厂需要进行原材料的采购，并且购置大型设备进行制造，大数据就可以看成最终数据产品的生产资料。

人工智能则可以与大型机械这样的生产力对应起来，看成加工原材料的工具流程或方法。生产的产品需要通过物流递送到用户手中，而云计算就提供了这样的供应链。这里的云计算不仅仅链接了工厂和用户，也链接了生产资料供应商与工厂，可以说是真正打通了上下游。

上文中提到的大数据与人工智能概念，算是狭义的。在实际工作中，做大数据的人往往也掌握人工智能技术，从事人工智能相关工作的人也离不开大数据。大数据和人工智能互不包含，当要给两者下定义的时候，得看我们是否在这个场合下同时谈论两者。当仅仅谈论大数据的时候，人工智能便是解决行业大数据场景下的具体技术而已。而当为了烘托人工智能的地位时，大数据便退化成为喂养人工智能的数据来源了。当两者在同一个场景下被提及，生产资料与生产力的提法或许比较形象有趣。

既然是为数据产品经理服务，那必谈数据。大数据这个词已经伴随我们好多年，若不是公司的大数据战略，又怎么会在行业内掀起一阵数据产品经理的旋风呢？所以，归根结底，数据产品经理这个岗位的设立得感恩于大数据概念的兴起。下面将介绍大数据的历史，以及其自身的发展。

5.1.1 大数据历史

1946年，世界上第一台通用电子计算机 ENIAC 诞生。参与建造 ENIAC 的科学家正是莫希利与埃克特，这两个人我们一会还会谈到。

与计算机同一时代的发明还有电话。不过电话相对于计算机发明较早，早在1876年，贝尔就发明了电话。

1948年，一位著名的科学家香农正在贝尔实验室工作，发明了奠定后来计算机存储以及互联网信息传输的理论，称之为香农定理。香农使用采样技术对语音模拟信号进行编码，从而形成数字信号，传输后再还原成模拟信号。由于01信号进行传输的时候高低电平几乎不会有翻转，因而保证了信息的完整，即便上面加载噪声，也能识别出高与低，而还原的清晰度则取决于采样的多与少，很好地解决了电话远距离传输存在的问题。可以说香农定理开辟了比特传输的新时代，为计算机存储与数据传输做出了贡献。

就在 1951 年，基于前人的成果，数据库诞生了。发明数据库的厂商是雷明顿兰德公司。这家公司原本是生产手枪的公司，这使得它与政府有很好的合作。二战后，政府对枪支的需求降低，并且开始转而专注于国内的经济的发展。进行经济发展就需要进行人口统计，当时的人口统计还仅依赖于卡片与纸袋，成千上万的录入人员坐在一起进行数据的录入与统计，于是雷明顿兰德公司就转而生产用于数据录入的打字机，并发明了 qwert 键盘（现在键盘的前身）。

1951 年时，国家统计数据需要依赖电子计算机进行，而电子计算机明显不能够使用纸带进行存储，这便孕育了数据库。由此便很好理解，为什么数据库最开始是关系型的，因为原来卡片上的数据就是填写在表格上的，很自然地便将同样格式的数据进行电子化。而就在数据库发明的这一年，美国就使用了电子计算机进行统计进而分析人口数据。为美国政府“服役”的这台计算机名叫 UNIVAC，发明人就是当年发明 ENIAC 的两位，而雷明顿兰德公司正是 UNIVAC 的资助者。

时间到了 19 世纪 60 年代，日本科学家梅棹忠夫（Tadao Umesao）发表《论信息产业》，提出信息化的概念。1969 年，互联网诞生了。其最开始是美国军方的一个项目，主要连接了 UCLA（加州大学洛杉矶分校）、STANFORD（斯坦福大学）、UCSB（加州大学圣芭芭拉分校）、UofUtah（犹他州立大学）四所学校的主机。

1970 年，关系型数据库才真正诞生，发明关系型数据库的人是埃德加·弗兰克·科德，他当时为 IBM 集团工作，也被称为关系型数据库之父。

1980 年，《第三次浪潮》（*the third wave*）出版，作者是阿尔温·托夫勒，这标志着信息革命已经被社会广泛认知。

1989 年，万维网 WWW 诞生了，蒂姆·伯纳斯·李发明了它。万维网和互联网的不同在于，它建立了一个公认的标准 HTML，使得人们可以发布自己的网页供大家检索信息。

1991 年，数据仓库的概念被提出，比尔·恩门撰写了《建立数据仓库》（*building the data warehouse*）一书，因此他也被称为数据仓库之父。

时间迈过 20 世纪来到 21 世纪，2003-2006 年间，以谷歌为首的一批互联网公司大展头角。Google 数据处理的三驾马车 GFS、MapReduce、BigTable 被相继提出，开创了真正的大数据时代。

2005 年开始, 大数据处理已经不再是谷歌这样的大公司的专利, Hadoop 诞生并被 Doug Cutting 所在的 Apache 基金会开源, 这使得无论发达国家还是发展中国家, 无论大公司还是小公司, 都能够使用大数据处理技术。至此, 大数据技术迈入了一个飞速发展的阶段。

5.1.2 自身发展

伴随着数据处理技术同时发展的, 自然也有和这些技术相生的数据。没有数据, 这些工具也派不上用场。

应该说, 最初数据是存储在纸张上的, 因而数据在没有计算机的年代就是被锁在政府档案馆中的一沓厚厚的文件袋。可以说有多少机关部门, 就有多少数据。

随着电子计算机技术的发展, 数据逐步被转移到存储介质中, 数据库、数据仓库都成为了数据的居所。企业也将自己曾经老套的账本和目录升级为了 CRM 与 ERP 系统, 这算是数据的另一个居所。可以说, 有多少企业就有多少数据。

再到后来, 电子计算机已经变成了便携的电脑和掌上设备, 人们也通过互联网相互链接, 产生数据的不仅仅是一个人, 而是一个网络。图片、音频、视频等多媒体数据, 传感器采集的数据, 以及互联网数据等多维度、多模态的数据都汇聚在一起。可以说有多少人就有多少数据, 不夸张地讲, 甚至可以说有多少设备就有多少数据。

而如今我们已经被环绕在“大物移云社”的环境中, 即大数据、物联网、移动、云计算和社交网络环境。这些我们看得见或看不见的环境无时无刻不在产生数据。根据 EMC 的估算, 到 2020 年, 这个地球大致会产生 44ZB 的数据, 要知道这个数据在 2013 年才只有 4.4ZB。什么是 ZB 呢? 1TB 等于 1024GB, 而 1ZB 等于 10 亿 TB。多么神奇的数字!

大数据润物细无声地潜入我们的生活, 影响我们的生活。我想也许有一天大数据的热度会消亡, 但是大数据开辟者带给我们的智慧启迪, 大数据先贤们带给我们的创新引领, 以及大数据技术带给社会的巨大进步不会消亡。

5.2 大数据的全球格局与中国面貌

信息革命自近现代兴起以来，至今已经走过了数十年，而以数据为代表的新一波科技浪潮却方兴未艾。尽管美国是信息技术的发端国家，也是创新能力很强的国家，但只要没有人阻碍全球化的进程，这样的科技就会在贸易中得以传播，被其他国家所袭承。

面对大数据的攻势，现在讨论的已经不是要不要做的问题了，而是怎么做的问题。我们将从政府这个层面来阐述各国对于大数据的扶持力度，进而详细介绍中国在这样的国际格局中是如何发展自己的大数据技术的。

5.2.1 全球格局

图 5-1 罗列了中国、美国、英国、欧盟、澳大利亚以及日本在政府层面对于大数据布局的相关政策。简单来看，政府发布的仅仅是促进大数据发展的政策文件，但是深入来看，大致可以看出这些政府关心的大数据产业究竟集中在哪些领域。

无论是企业的战略还是个人的成长，都应该顺应时代的潮流，如何把握时代的脉搏就是值得我们探讨的问题。对此，政府的政策性指引给了我们发现趋势的线索。







	“十二五”新兴产业发展规划提出大数据 2012.7		《促进信息消费意见》 提出构建大数据产业链 2013.8		工信部发布 大数据白皮书 2014.5		国务院《促进大数据发展行动纲要》 2015.8		国家发改委发布 大数据发展重大工程 2016.1		
	《大数据研究 和发展计划》 2012.3		棱镜门 曝光 2007-2013		《大数据：抓住机遇， 守护价值》 2012.3		开放“我的学生 数据” 2014		政府数据全开放 Data.gov.us 2015		
	开放式数据 研究院ODI 2012		向大数据技术 注资1.89亿英镑 2013.1		与李嘉诚基金会进行 医疗大数据研究 2013.5		《把握数据带来的机遇： 英国数据能力战略》 2013.10				
	发布 《数字化路线图》 2013.2		发布《法国政府大 数据五项支持计划》 2013.7		将大数据列入 地平线2020计划 2014						
	《公共服务信息与通 信技术战略12-15》 2012.10		发布公共服 务大数据政策 2013.8								
	建设活力 ICT日本 2012.7		医疗、交通、 新媒体大数据 2013		创建尖端 IT国家宣言 2013.6		大数据个人 信息保护法 2014.6				

图 5-1 全球各国与地区关于大数据的举措

首先,数据开放是根本。从英国政府 2012 年开始建立开放式数据研究院(ODI, Open Data Institute),到美国政府建立数据开放网站 data.gov,再到中国在北京与上海进行政府数据开放来看,足以表明政府对于大数据的态度。公开与开放是创新的第一步,将数据向社会开放,可以使用了民众的大众智慧来对数据进行创新,这对于政府来说何乐而不为呢?

其次,重点行业优先投入。在各国的行动计划中,医疗、工业都是优先发展的行业。医疗关乎民生,而工业则是实业,也是一个国家立国的根本。英国和日本都建立了相应的机制在医疗行业进行大数据投入。而对于实业来说,德国则提出了“工业 4.0”,中国也提出“中国制造 2025”战略,即采用数字、智能制造等技术进行绿色环保的工业革命,一次性根除前三次工业革命给社会带来的顽疾。除了医疗与工业,交通以及新媒体等领域也均有涉及。

最后,大数据隐私很重要。数据的开放固然能够给社会带来高效与便捷,然而数据公开就意味着我们需要拿自己的隐私作为交换,这不仅仅是技术层面的事情,更是道德与伦理层面的事。于是数据的公开就没有想象中那么顺利,这不再是政府一厢情愿的事情,而需要每个民众都参与其中,所以出台的政策想民众之所想就显得尤为重要。

有两个视频可以推荐给大家,是英国 BBC 公司拍摄的关于数据的纪录片,分为上下两集,在腾讯视频搜索“汉娜的故事”可以找到。这是一部英语旁白中文字幕的记录片,但愿能够加深你对数据的感知。

5.2.2 中国面貌

将视角切回国内,我们就会发现,从“互联网+”战略到“中国制造 2025”,每一处都有数据的参与。互联网本身就是数据的直接产生者与消费者,而数据又是用来驱动制造业进行绿色制造与柔性制造的一把利器,于是中国自然不会放松对数据的关注。

中国对于数据的布局主要集中在四个区域,分别是京津冀、长三角、珠三角以及以贵州为首的中西部。以北上广为代表的京津冀、长三角、珠三角等地区发展大数据并不奇怪,经济基础雄厚是其最大的优势。

贵州作为一个相对偏远的地区能够发展起来，可以说没有一定的战略眼光是不行的，但贵州的兴起同时也要归功于其地理位置和气候条件。

众所周知，在大数据之前有一个概念叫作云计算，而大多数人对于云计算的理解就是大型机房里摆放了众多的机器。这些机器堆积在一起需要散热，于是需要建立空调系统来保养这些机器。贵州这个地方的气候对于需要冷却以及降温的机房而言比较合适，且地质活动相对稳定，也不会有地震等因素造成数据的物理损失。由于是一个相对偏远地区，因而人力成本和税收相对便宜，这也降低了企业发展云计算和大数据的成本。这也是谷歌为什么会把数据中心建在湖泊和大海边，更有甚者将数据中心建在北极的原因了。

一眼万年，各国对大数据的布局是历史进程中的一小步，类似大数据的概念还会层出不穷，终了是要能够使用不断发展的技术为人类所用。任何真心付出的，坚定投入的，必将会有回报。正如李叔同在《晚晴集》中说的那样：“念念不忘，必有回响”。

5.2.3 行业概览

应用大数据的行业繁多，几乎每个行业都可以多少与之沾上边。这里我们将对大数据行业做一个快速概览，并在接下来的章节中为大家详细介绍大数据在各个行业中的应用点。

作为前提，有两点需要交待清楚。第一点是，当我们在讨论大数据的行业应用时，难免会与人工智能在该行业的应用有所重合。第二点是，很多行业都在炒作大数据，有一些是真正成熟的应用，有一些则属于尚未成熟的应用。怎么区别两者呢？有一种粗糙但简单的方法。如果一个应用是由企业自觉发起并进行商业探索的，且这些大数据应用便是使得企业成本降低或用户体验改进的，那么这样的大数据应用便是成熟且可以大范围推广的，毕竟企业以盈利为目标。如果一个行业的大数据应用是靠组织出台规定而兴起的一种“时尚”，大家纷至沓来都想分一杯羹，那么这样的大数据应用则多少有些不成熟，市场还需培育。在政策改变或风潮退却后，这样的应用往往难以具备内生动力，难以维系。

对待成熟与非成熟的大数据应用，我们的态度也应有所不同。对于成熟的应

用，我们进入这个行业只需要采用跟随战术切入即可，然后在细分领域或者新兴市场打开局面。而对于非成熟的大数据应用则更多需要开拓与探索。尽管会有人说非成熟的应用前景光明，但你很难说这些人是不是出于某种利益因素才与你画饼，在衡量专家的观点时，我们需要区别什么是事实信息，什么是个人看法。事实信息是第一源信息，可以用来作为自己的判断标准，而别人的看法只能作为参考。即使光明的前景是必然，现在也并不是最合适的时机，过程中需要承担巨大风险，毕竟收益与风险成正比。

图 5-2 大致描述了若干与大数据紧密相关的行业，后面的几节将详细论述，这里不妨让我们按照一个新的分类方式来介绍大数据在行业中的应用。十八大报告中提到了“五位一体”的概念，其中的五位分别指的是经济建设、政治建设、文化建设、社会建设与生态文明建设。下面就按照这五个方面对大数据与各行业进行介绍。



图 5-2 大数据与各行各业

经济建设

人们对于国家经济最粗浅的理解就是 GDP，按照支出法进行核算的 GDP 包括了居民消费、企业投资、政府购买以及净出口（出口-进口）。从这个意义上来看，经济关注了金融、消费以及制造业领域。因此大数据服务于经济则主要着力于大数据金融、大数据零售以及大数据制造。

对于大数据金融来说，其主要关注点在于风控。对于大数据零售来说，其重点在于精准营销。而对于大数据制造来说，考虑更多的则是如何利用数据来驱动制造业的改革，这也比较符合国家提出的“中国制造 2025”战略。

政治建设

姑且把法治建设归为这一类，这样大数据治理便顺理成章地可以成为辅助政府进行政治建设的帮手。如利用大数据进行舆情的关注，进行案件的辅助判别等。

文化建设

文化领域包括了教育、体育、旅游、医疗等方面。对于教育大数据，其关注的点主要在于帮助学生成长方面，这样的成长有学业的成长也有综合素质的成长。对于体育领域，大数据能够发挥的威力在于记录训练的过程，进而更好地指导训练。大数据对于旅游则更像是营销，能够帮助游客解决旅游过程中“吃住游购娱”的问题，除此之外旅游大数据还有助于满足政府部门的监管需求。而医疗大数据则更偏重于数据辅助的精准医疗方案。

社会建设

社会层面的大数据应用要数社管、智慧城市等概念最为人瞩目了。交通亦是智慧城市领域的一个应用。大数据可以帮助进行路况预测与交通疏导，以及城市规划等。对于社管来说，原本的定义应该包含了社会治安、教育等领域，但这里为了赋予每个领域一个字面意义上的归宿，于是社管的定义就被人为地缩小了。

生态文明建设

农业应该算与生态沾边了吧。大数据助力农业看似是件新鲜事，在中国，尽管东北一些地方的现代化农业已经卓有成效，但毕竟大片的耕地还是小集体种植，这样大数据农业的应用面可能就没有那么广了。利用大数据可以帮助农业预测天气、产量，指导种植过程。技术一旦推广并普及开来，应该算是功劳一件。

再次审视每一个领域，所有大数据与该领域的结合都是使用大数据助力其增效，或者使用大数据帮助其业务降低成本，最终的目的都是辅助业务，而业务模式和流程创新只是手段。任何的应用出发点都不应该偏离这个领域业务的初衷，所谓“不忘初心”大概如此吧。

5.3 大数据+“治理与交通”

之所以把大数据和治理与交通放在一起，是因为它们都属于智慧城市的一部分。在很多企业中，智慧城市业务基本上也会涵盖这两部分。由于是政府出资建设，因此只要具备相应的资质（如技术资质，保密资质等）再加以一定运作，就可以锁定目标。下面分两部分分别介绍治理与交通。

5.3.1 治理

对于政府治理来说，大数据大致可以有下面五个用武之地。

舆情监控

随着社交网络的普及，以前只有权威人士才可以发声的媒体（如电视、报纸等）逐渐被大众性媒体所取代。只要你愿意，就可以注册一个微博或微信，或者像很多自媒体人一样，开通一个头条号或者微信公众号。这同时也造成了一个问题——舆论变得更加熙熙攘攘，难以管控。通过大数据对网络文本甚至语音进行分析，可以了解民意趋势，也可以粉碎谣言并找出源头造谣者。这些应用都可以称为舆情监控。

在舆情监控领域，需要掌握的技术有爬虫技术、自然语言处理（Nature Language Process, NLP）技术等。使用爬虫技术是为了抓取网络媒体上的文本信息，从而使舆情监控有可以生火做饭的“米”。而自然语言处理技术则是将米饭煮熟的火。抓取下来的文本从格式上看较为杂乱，一般被称为半结构化数据。从细节处看，自然语言处理技术对文本进行分词、实体识别；从整个文本看，自然语言处理技术又可以对其进行情感判别、文本分类等。通过这样一系列技术，就可以搭建出一个粗糙的舆情监控系统。当然，针对具体的业务，不同的舆情监控系统略有区别，这取决于业务的理解与用户的需求，但其核心技术离不开上述内容。

监控视频分析

除了手机里的自拍照片，其实我们也悄无声息地被别人拍摄了许多照片。例如在高架道口行车，步入超市购物，小区散步，机场候机，办理证件等，城市中心联网的摄录机以及各种图像采集设备已经将我们的身影记录下来。卞之琳有诗云：“你站在桥上看风景，看风景人在楼上看你”，便有此番意味。

这些图片和视频数据也是非常有意义的。就拿天网监控视频来说，既可以用来追踪小偷，也可以帮助找回被拐儿童。当然，人工智能与大数据技术还未智能到电影中那样，或者即便已经技术昌明但是却未能大规模普及。图像中人脸的动态识别，以及模糊图像的增强显示都是需要克服的困难。但是，一旦这些技术广泛应用开来，这将是视频分析技术迈出的坚实一步。现在比较成熟的是交通卡口的车牌识别，安防领域的人脸识别也是最为常态化的应用之一。

监控视频的分析与图形图像分析密不可分。现在的视频是由一幅幅图像组合而成的，对视频的分析可以拆解为对图像的分析。因此其背后的核心技术主要是图像处理技术。随着深度神经网络的发展，以及海量数据的标注，人脸识别、物体识别、动作识别已经在多个领域中慢慢普及开来。

群体性事件预测

当我们说群体性事件的时候，一般指物理环境中聚集众多人而有可能引发的社会不稳定现象或者因为大规模人口聚集而造成的伤亡事件。举个例子，2014年年底的跨年焰火晚会，上海外滩爆发了踩踏事件，这就是一个群体性事件。那么如何才能预测此类事件呢？一种简单粗暴的手段就是在各个地方装上摄像头，人为去监控或者让机器去分析。人为监控不可取，机器分析技术虽可行，但摄像头覆盖范围有限且相关摄像头还有待联通。

一个别出心裁的方案是，可以通过运营商的手机信号数据进行预警，大致了解人们都处于什么位置，并在地图上标出每个人的位置，于是就可以看到地图上的热点图。有些像卫星云图一样，通过前几个时间点的人群位置移动，就可以大致预测人们都在向哪些地方汇集，哪些地方可能发生群体性事件。

这种可视化的预警方式有些类似于灾害天气的预报。相似的原理借鉴到大数据治理上可以达到相同的效果。

宏观经济数据监控

由于大数据是伴随着电子计算机和信息技术以及互联网技术的兴起而扩散开来的，因此我们很容易先入为主地认为机器采集的数据才能构成大数据。数据的来源很重要，但更为重要的是在现有的采集手段下如何进行数据分析。

我们在介绍大数据的历史时，就说过了人口统计对于国家的重要性，这里的

人口数量实际上就是国家宏观经济关心的一个指标。除此之外，各地的消费品价格，每个地区公司的数目，公司雇佣的员工数目，各个地方的招商引资情况等都形成了宏观经济的数据报送。无论这些数据是以计算机系统的形式记录在案，还是以纸质的形式定期呈报，其本质都是为了国家的宏观决策而制定。很难说这些数据和电商网站的数据比起来是否算大，也很难简单使用多少字节来度量它们，但是它们的价值却并不逊色于电商大数据。一个更有趣的例子是，可以使用电商数据作为社会数据的采集神经末梢，因而也可以指导政府宏观政策的制定，电子科技大学的周涛教授团队就在这个方面获得了许多成就，感兴趣的可以搜索一下“新经济指数”。与之相似的，淘宝也有类似的指数 aSPI 与 aEDI。

民生应用创新

政府能够主导的研发毕竟有限，其主要功能还是协调与统筹。但是通过将数据开放给社会，激发社会中的创新、创业力量，让社会中的有志之士基于这些数据进行创新，这样等于激发了整个社会参与数据应用的潜能。例如，创建一个找停车场的 App，或是打车的应用，总之人民群众的想象力是无穷的。上海在这方面走在了全国前列，举办了几届数据开放大赛 SODA 就是最好的印证，同时取得了不错的社会效应。也有一些应用最终落地并被政府采用，成为大众服务的最终产品。

5.3.2 交通

交通出行是人们最常见的民生活活动，因而在政府治理领域尤为被重视，也是智慧城市诸多应用的突破口、切入点与发力点。大致说来，交通大数据可以应用在下面四个方向。

车流密度监控

该项技术有些类似政府治理中的群体性事件预测，但对于交通来说，更多的是通过车量在地图上（主要是道路上，而非居民区）的可视化情况，显示各条道路的车流密度。在很多城市可以看到道路旁有一块电子大屏幕，上面显示的城市道路简图中会将有些道路标注为红色，有些标注为绿色，被标记为红色表示该道路拥堵，而绿色则代表该道路畅通。该应用利用的就是车流密度监控技术，将道路车流信息展示给社会大众，让大家自主决策。

交通拥堵疏导

2016 年 10 月，阿里巴巴集团在云栖大会上发布了“杭州城市大脑”，可以通过数据了解道路交通状况，并且可以与信号灯控制系统联动（重点在这），以调节信号灯的红绿状态，使得道路通行状况最优。这算是智慧交通的一件大事了，想想曾经在空无一人的道路上也需要等红灯的两难尴尬，阿里的这套系统可以帮助化解，这也算是上一个实时车流监控的升级版了。

公交线路优化

上述应用都是实时的，当数据沉淀下来之后，大致就可以了解一些人的固定行为模式。譬如我每天早上都从“霸都”合肥的东城开车前往西城工作，在了解了大规模用户的行动模式之后，就可以按照这个模式来设定公交线路。哪些线路流量大需要增加线路数量，哪些线路上需要设置更多的车站，大数据可以辅助城市规划人员，成为其工作助手，更好地进行城市的数字规划，弥补以往经验性规划的不足。

城市功能区划分

借助大数据除了可以了解人们的行为模式，还可以了解人们的生活工作区域，例如合肥东城就是一个生活聚集区，而西城有高新区，显然就是一个工作区域。在上班时，人流从东到西呈现顺差，而下班时间从东到西则呈现逆差。也许你会觉得两边对流人口一样多，那不过是车多路窄的假想罢了。通过对城市功能区的认识，可以指导单位在审批新项目的时候，统筹考量经济目标之外的社会效应，以达到城市有效运营的终极目的。

5.4 大数据+“零售与金融”

有一种说法是，按照当下产品类型容易变现的程度来进行排序的话，首当其冲的是电商，紧随其后的是社交，然后才是内容。在与大数据相结合的行业中，也有两个行业离金钱很近，或者说较为容易变现，分别是零售和金融行业。

5.4.1 零售

零售的概念大家应该不陌生，从 24 小时便利店“711”到街边的水果店，从

网络订餐平台到苹果的线下体验店，这些都是零售的形式。这些零售的实例中往往有两个主体，一个是商品经营者，另一个是商品消费者。从其交易活动中我们大致可以总结出两个特点：一是单次交易商品数量少，与零售相对应的是批发，批发的单次交易商品数量更多；二是交易的次数较为频繁，商场与水果店可能每天都会被访问，而订餐平台的复购率也相对较高。

以现在比较流行的水果专卖店为例，我们不妨看一看零售的过程中需要用到哪些推销手段。首先是广告，同类的水果店很多，要想在众多竞争者中脱颖而出，就需要通过广告来宣传自己的经营理念。所有的广告都指向一个目标，吸引顾客购买，从这一点来说，所有的广告都是文案层面的 PK。其次是推荐，通过推荐，帮助顾客找到自己想要购买的水果，就需要看店员的本事了。店员可以根据性别为其推荐美颜应季水果，也可以根据年龄推荐养生果品，还可以根据消费层次推荐打折或高端水果，甚至可以根据消费目的（如看望病人）推荐果篮等。通过推荐，不仅可以帮助顾客更加便捷地找到自己想要购买的商品，还可以起到提高客单价的作用。

无论是广告，还是推荐，都可以搬到互联网上。一个在线零售企业的大数据策略也必然包含这两者。图 5-3 是我找到的两个示例。



图 5-3 推荐系统与计算广告示例

左图是推荐系统示例，截屏自亚马逊的 App 页面，你可以在任何电子商务网站上找到这部分，叫作“猜你喜欢”。那段时间我正在关注教育大数据和数据分析方面的内容，也采购了相关的书籍，于是便有了图中的推荐结果。右图是从今日头条截取的图片，虚线框出来的就是计算广告的结果，换一种媒体人或营销人更加熟悉的称谓，叫作“精准营销”。正巧那段时间我正在搜索周末度假的地方，于是头条就为我推送了这则广告。

推荐系统

对于推荐系统来说，其大数据后台的运作机理大致有两种，可以概括为“物以类聚，人以群分”。一种是把和你类似的人喜欢的东西推荐给你，另外一种是把与你喜欢的物品相似的物品推荐给你。

推荐系统已经成为了一个研究领域，算法也层出不穷。在学术界，有一个推荐系统领域的国际会议 RecSys，每年在该会议上，专家学者们都会发表大量的最新研究成果，力求提升推荐算法，介绍新应用。随着互联网产品的多元化，推荐的内容也越来越多样化。从最开始的商品推荐，扩展到现在的店铺地理位置推荐、音乐推荐、新闻推荐以及好友推荐等。而这背后的算法也因推荐内容的不同而有所区别。项亮老师（我的师兄），曾是美国奈飞（Netflix）推荐系统比赛的第二名，是推荐系统方面的专家与先行者，著有《推荐系统实践》一书，书中详细介绍了推荐系统各种算法的机理与应用。

计算广告

计算广告或精准营销的过程略复杂。简单来说，就是当你在手机上看新闻时，大数据知道你是谁，知道你的偏好，于是在这个新闻 App 的横幅、底栏或内容中夹杂一些你可能会点击的广告。深入来看，可以使用图 5-4 来解释。

在计算广告中，有三伙人，一伙称为广告主（DSP，Demand-side Platform），就是想要做广告的人，这些广告主大多是企业，涉及各行各业。第二伙是媒体，也称为供应商（SSP，Supply-side Platform），可以是一些网站，或者手机 App 等，用于提供广告位。实际上，并非所有商家都采用计算广告的方式来进行运作，有的（譬如横幅广告）可能和公交站台广告一样，一次性承包给 DSP。第三伙是连接器（ADX，Ads Exchange），有点像介绍人，把广告主介绍给媒体，或者把媒体

介绍给广告主。他们会把媒体资源集中后握在手上，也会把要做广告的人聚拢起来，然后在这两者之间进行匹配。当然，有的时候 ADX 可以由 DSP 的人来充当，有的时候媒体也会自己充当 ADX，此处，姑且让我们认为 ADX 是独立的第三方。



图 5-4 计算广告流程

ADX 的技术含量是什么呢？首先他们会研发一套工具，媒体在开发 App 的时候会使用这样的工具去生成可以展示广告的位置。这个工具还可以收集使用这个 App 的用户信息，根据这些信息知道用户的喜好。那么用户的喜好是怎么存储的呢？中间商会建立一个数据库来存储每个用户（实际上是每台设备的编号）的喜好（用户画像），这个库就是数据管理平台（DMP，Data Management Platform）。有了 DMP，计算广告就可以实现了。

下面通过一个场景来看看计算广告的大致流程，当刷新今日头条的页面时，后台是如何展示某个广告的呢？首先，刷新页面的时候，展示广告的地方会使用工具把设备编号传回后台，后台通过设备编号在 DMP 中查询这个编号的偏好，查询完毕后要了解现在打算做广告的人有哪些，把要展示的广告和用户的兴趣结合起来，估计持有这台设备的用户点击某个广告的可能性有多大，最终选择可能性最大的推送到你的手机上。整个过程，从刷新到展示只需要不到 100 毫秒。这就是计算广告。

刘鹏老师的《计算广告》中对计算广告个中丰富内涵有更为全面的解读，从技

术到策略，从历史演变到商业模式，无所不谈，是难得的计算广告优秀入门书籍。

5.4.2 金融

大数据金融是大数据直接作用于金钱的最典型行业案例。在介绍金融领域的大数据作为时，我们要明白两个时间、三个对象以及两个概念。

两个时间

说到金融，必谈风险。所以金融大数据就是围绕风险管理展开的。这两个方向就是按照授信时间的前后而划分的。

在授信前，大数据助力决策支持，对用户进行征信，看看到底应不应该授信给这个用户，如果授信，到底授信多少金额。

在授信后，大数据助力风险管理，譬如说能否即时还款，是否存在欺诈，是否存在洗钱等违法违规行为。

三个对象

如果将大数据助力金融从对象的角度进行划分，可以分为个人、企业、市场三个角度。

对于个人来说，可以进行精准营销（毕竟金融连着消费）、风险定价（决定授信额度）以及反欺诈（授信后还款与否）。

对于企业来说，可以进行风险监控（股价波动评估、企业风险评级等）、风险定价（帮助企业借贷和融资）以及反欺诈。

对于市场来说，主要是进行量化交易（即计算机自动根据收集的信息决定什么时候买哪只股票卖哪只股票）以及财富组合管理（类似于余额宝等投资产品组合）。

两个概念

这里要介绍的两个概念分别是 FinTech 与 BlockChain。

FinTech 是 Finance Technology 的缩写，即金融科技。大数据金融代表技术视角，这往往是 IT 界人士提出的，而对金融从业者来说就是 FinTech，符合金融精英们的高端气质。关于 FinTech 的具体领域，上面提到的基本都算，除此之外还

有一些关于金融商业模式的内容，要想了解详情可以从网上找一篇关于 FinTech 领域的文章阅读，此处不再赘述。

BlockChain 是区块链，是目前最火热的金融科技概念之一。区块链可以这样理解，为了证明某笔交易确实发生，使用计算机数学算法替代第三方信任保证。

例如，我们从淘宝网上买东西，给店家打了一次钱，店家不会重复要，因为我们先转给了支付宝，支付宝就是第三方信任担保。传统的银行和支付宝起到的作用是一样的，都有担保的功能。我们之所以信任它们，一方面是相信品牌的力量，另一方面是相信资金的力量，政府政策中要求它们缴纳的准备金足够多。

而区块链是什么呢？是使用互联网特性（信息公开共享特性）来记录信息，保证大家不会赖账。还是以向店家付款为例，如果不再通过支付宝而是使用区块链，那么转账给店家时涉及的信息（包括转账人、收款人、款项、钱数、时间等）就会被记录在互联网中，并可以被所有人知道，这样一来担保就不只来自于一个独立的信任第三方，而是整个互联网群体，谁也赖不了帐，因为每个人都可以作证。

区块链有风险吗？当然有，如果对方和除你之外的所有人串通起来，这个系统就等于被破解了，但是这件事情何其困难，如果计算概率，比银行或支付宝记账失误的概率可能还要小，另外，区块链上的信息也几乎篡改不了！

还有一种更简单的方式可以解释区块链。以前的传统交易往往找一个德高望重的人来见证，所以成为第三方信任担保的交易。而区块链技术是让两个人的交易在众目睽睽之下进行，每个人都可以为交易证明。

5.5 大数据+“体育与教育”

5.5.1 体育

大数据助力体育？不知道大家会想到什么。是央视 CCTV5 关于世界杯、欧洲杯的数据统计，还是懂球帝（一款球迷社交 App）上关于球迷的数据运营。尽管 CCTV5 与懂球帝都与体育和大数据相关，但我今天想谈论的并不是这个。

如果你阅读了大量的大数据文献，并且了解大数据在各个行业中的应用，你一定会对这两个词有过思考：大数据××与××大数据。以体育作为例子是最适合不过的了。之所以选择体育是因为体育的大数据相关技术应用并非那么成熟与深入，因而这两个概念在这个领域比较容易分开。那么“体育大数据”是做什么的呢？而“大数据体育”又是什么？为了便于理解，我们不妨将这两个概念稍微“翻译”一下。

“体育大数据”可以“翻译”为“体育相关业务的大数据展示”，而“大数据体育”则可以“翻译”为“大数据技术助力体育业务”。这样就可以清楚知道这两个概念在本质上的区别了。对于体育大数据来说，成果是信息展示图，而对于大数据体育来说，成果是使用技术使得体育业务降本增效。下面给出一个大数据体育的例子。

训练数据分析是大数据体育最典型的应用场景。还记得我们在前面提到的“大物移云社”的概念吗？“物”就是物联网，各种穿戴设备就是物联网的组成部分。设想一下，每个运动员都带上腕表，他们的心率、步数、睡眠、位置等都可以被采集。那么我们可以基于这些训练数据做什么呢？大致有以下四点。

体能建模

体能建模主要是针对运动员进行体能上的建模，我们通常说的爆发性选手、耐力型选手便是对一个人体能的描述。可穿戴设备可以记录运动量、血压、脉搏等数据，可以很好地对一个人的体能进行描述。运动员的生理水平随着比赛时间的增加如何变化，在大赛前的竞技状态如何，这些指标通过对一个人的数据进行长期采集和建模可以得出。

转会评估

在足球俱乐部中经常会发生转会的情况，往往转会费高达千万美元。俱乐部作为一家公司，在考虑投入重金引进球员的时候，免不了要评估这个球员给俱乐部带来的投入与产出比。除了考虑明星球员作为个人IP能够给俱乐部带来的流量价值外，球员从战术水平上表现如何，能否与现有俱乐部球员间互补也是一大考量因素。大数据可以将球员的个人体能模型与俱乐部内部的体能模型进行比较，从而得出一个评估的结论。

教学管理

中国各个地方涌现出了许多体育学校，也有很多家长愿意把孩子送到这样的学校中学习体育。通过大数据可以对日常的体育教学日程进行管理，并帮助家长了解孩子在校的表现。这里大数据的用武之地更像是在体育这个垂直领域的教育应用，如教学视频、家长信息推送等。

战术分析

在看足球的时候，我们时常会听到一些关于打法的描述，如“4-3-3”的进攻阵型，又如“5-3-2”的防守反击阵型。俗话说“兵无常势，水无常形”，实际比赛中的阵型不是一成不变的，大多数情况下俱乐部会通过录像回放来分析战术，而现在大数据可以通过分析录像中对方与我方球员在赛场中的跑动得出实际比赛中阵型的变化。

5.5.2 教育

做教育大数据的人，大致可以分为两拨，其中一拨是从教育行业成长起来的，我们称之为教育专家；另外一拨则是从IT领域转型而来的，我们称之为IT专家。

在大数据应用于教育这个行业来说，有两个死锁状态难以破解，一个是人，一个是数据。其实这样的死锁状态在各个行业的大数据实践中都存在。对于人来说，两个行业的人相互等待，不清楚对方的初衷。现在看来，教育大数据实践，妄动占大多数，不是太浅显（纯属数据统计报表），就是太跑题（曲高和寡的研究，难以落地，难以真正服务于教育）。

对于数据来说，一方面是拥有数据的人会质疑挖掘数据价值的人是否具有专业水准，而挖掘价值的人则会希望先拿到数据再看可以做什么事情。就这样，“人”与“数据”两个死锁，让教育这个行业以及类似的行业陷入发展大数据的尴尬境地。

那么下面就来看一看大数据真正助力教育的场景下所涉及的三个维度与五个对象。

三个维度

三个维度分别是平台、内容与工具。

对于平台来说，教育信息化主要是指建立政府的云平台（资源与空间平台）、MOOC 平台（如 coursera、网易云课堂）、SPOC 平台（线上与线下的互动的学习平台）。大数据在这些平台上的作为，可以是使用资源分析工具对资源数据进行结构化，基于平台用户网站使用行为对用户进行数字化营销（挽留课程流失学生），当然，还可以是基于全系列的教育产品进行的学生与教师行为分析，这些分析很难说有什么直接的作用，但是至少给大数据助力教育提供了一些思考点。

对于内容来说，主要有多媒体资源、教学素材以及题库。除去上面谈到的资源内容结构化与精细化标注之外，还有基于题库做题的自适应学习。

对于工具来说，主要是从教、学、考、评、管、研六个角度切入，类似于背单词、英语听说练口语、走班排课系统以及智能批阅系统等。这些之所以称为工具是因为其可以基于自身产品进行小闭环式的迭代，并不需要大量的外部数据。

五个对象

教育中涉及的五个对象分别是教育主管部门领导、校长、教师、家长和学生。他们虽然都围绕教育发生联系，但是各自诉求不同，因此要想使用大数据助力教育，需要分角色对待。

对于教育主管部门领导来说，我们要透过需要看动机。即便是主管部门，不同位置上的领导动机也不同，从最高位置的领导到基层的领导，关注的方面各不相同。

对于校长来说，他关注的点可以是校际趋势、学科均衡与教师评价。在产品中需要着重强调对比。

对于教师来说，教学改进与学情分析都是他们的关注点。相对于领导而言，教师的诉求更加实际，大数据能够帮助教学工作或帮助其省时省力的点，才是教师真正欢迎的。

对于家长来说，学情诊断推送与人文关怀（教育咨询、心理教育等）或许是一个不错的关注点。尽管学生是学习的主体，但在高等教育之前的阶段，家长才是真正的付费群体，满足他们的需求势在必行。

对于学生来说，围绕国家的教育改革要求进行生涯规划、选科规划、报考决

策、个性化辅导都可能是产品中可以涉及的点。如果非要把提高分数这件事加进来自然也顺理成章。

这样的框架并非穷尽所有的功能与产品方向，但在考虑产品方向的时候会提供一个思路。当然这样的框架并非一成不变，你可以建立自己的，或者在此基础上扩充与完善。

5.6 大数据+“医疗与旅游”

可能有人会问，为什么将医疗与旅游放在同一节中介绍呢？

首先，它们都具有较强的共同服务属性。其次，医疗与旅游二者共享着大规模的用户群体。

这些理由促使我将医疗和旅游两个行业归纳到一个小节中进行描述，介绍大数据在这两个行业中的分别应用。

5.6.1 医疗

医疗除与旅游行业有共通点之外，与教育行业也有很多的相似点。首先，两个行业中的客户在进入这个行业的时候都对未来有预期的收益，而这个行业的从业者往往却很难满足他们的期望，这使得这两个行业的平均客户满意度相较于其他行业较低。其次，这两个行业都是经验行业，难以被标准化，并且稍微一些的经验差异就会造成结果的很大不同。

总体来说，大数据助力医疗大致有四个主要用途。

互联网医疗

想必大家一定听说过互联网+医疗，离我们生活最近的应用就是网上挂号。当然并非仅仅如此，放宽概念范围应该称为便民惠民应用，如公共疾控、药品定价等。它实际上是将生活中原本需要线下进行的工作搬到了线上。

医疗大数据

当我们入院之后，会有很多数据沉淀在医疗机构。健康状况、用药状况，甚

至刷卡记录等，拥有这些数据的医疗机构可以做什么呢？大致来说，可以分级诊疗与个性化药事服务。所谓分级诊疗就是按照疾病的轻、重、缓、急以及疑、难、繁、杂程度进行分级，不同级别的医疗机构承担不同疾病的治疗。而个性化药事服务就是针对各个人的身体情况，使用海量的数据和固化的名医经验进行治疗方案的推荐。这样的大数据应用实际上是将人的经验数据化，并使用数据化的经验提升服务效率与个性化程度。

医疗管理

医疗管理有些类似于城市管理，主要是针对于医疗监控视频进行分析。这里就不赘述了。除此之外还有一些类似于医疗内部管理系统的大数据分析应用，如果把医院看成一家企业的话，其本质与 ERP 的数据挖掘相同。

医学研究

这是一个前瞻的研究领域，只有高校、科研机构与大企业的研发院所才有可能涉及，譬如医疗影像分析、结构化电子病历、医学知识图谱等领域。国际知名巨头 IBM 早年启动的 Watson 计划中就包含智能医学影像分析项目，除此之外的中国“数字肺”项目也开启了医学影像大数据辅助诊断平台，其目的是进行肺部疾病的筛查。

5.6.2 旅游

旅游服务这种第三产业主要是围绕人们的需求展开的，概括来说就是“吃、住、行、游、购、娱”。我们可以用三阶段、四对象来概括。

三阶段

三个阶段是按照用户介入旅游的时间节点进行划分的，可以分为旅游前、旅游中与旅游后。游客在这三个阶段的关注点以及商家可以在三个阶段开展的活动是不同的，因而大数据会在这三个阶段有不同的表现。

对于旅游前来说，主要侧重点是景区人流预测与旅游套餐推荐。旅游主管部门需要对高峰和平峰客流进行预测，特别要对高峰时期各个时间段的人流情况有所了解，以便安排运力、安保以及服务，因而景区人流的预测就显得特别重要。

传统的景区人流预测主要基于景区智能设施的统计以及视频数据的分析,例如进入景区的计数围栏,以及关键路口的人流计数等,这些都属于大数据的统计方式,除此之外还可以使用手机基站的数据来进行景区流量预测,这样的预测相对于传统的预测来说,更为及时,且更为真实可信。

对于旅游中来说,主要是人流监控、客源分析、出行方式识别、驻留时间分析、团散比分析、多景点关联分析以及热门线路挖掘。

对于旅游后来说,则是游客情感分析与舆情监控。如果把旅游服务看成一个产品,从互联网领域增长黑客的角度看,做广告只能算是拉新,客情关系维护才是起到提升复购率作用的关键。

四对象

与之前介绍教育行业时类似,当我们提到对象的时候,往往指的是这个行业中的相关人群。在旅游行业中,他们分别是主管部门、景区、酒店等商家以及游客。

对于主管部门来说,他们关注的是战略规划决策与景区监管,大数据具体可以实现客观三方体验指数分析、景区人流监控与舆情分析。具体的大数据产品可以是景区决策中心的大屏幕与舆情分析系统等。

对于酒店来说,精准营销是关键,那么大数据的作用就变成了个性化推荐。大数据可以通过我们在计算广告一节中介绍的用户画像系统进行定向投放,以最小的成本获得最有价值的客户。

对于景区来说,如何提升游客价值与体验至关重要,于是类似于流量预警以及景点关联分析这样的方向就可以作为大数据的抓手。景区主要提供交通、安保等基础服务设施,寻找高效的运营游客方法,通过流量预警和景点关联分析为景区运营企业找到一个保障日常运营和提升游客体验的新思路。

对于游客来说,主要内容是产品推荐与信息获取,展开来说有旅游套餐、景区信息获取与旅游社交(如驴妈妈)等。

我们要面向不同的对象开发不同的大数据应用,要相信市场与舞台足够广阔。

5.7 大数据+“农业与制造”

把农业与制造业放在一起介绍再合适不过了。作为人类发端的助手，农业为数十万年的人类历史带来了能量，而工业则加速了人类在这颗蓝色星球上的崛起。我们常称农业为第一产业，工业为第二产业。接下来就谈谈大数据在第一、第二产业中是如何分别发挥它的作用的。

5.7.1 农业

对于农业来说，一切都是围绕农业生产而展开的。大数据可以进行数据采集，以前的经验种植需要农民人为去判断什么时候才是种植的最好时机，而现在这一切都在信息化时代和传感器革命后得到了解决。有人说互联网社会或人工智能社会的这些发明就是人类感官的延伸，这个说法没有问题。对于农业来说，我们使用传感器可以测量土壤酸碱度、气候、空气质量、作物成熟度，同时还可以使用信息系统记录设备使用情况与劳动力成本。

获得数据只是第一步，利用这些数据，我们可以做什么呢？主要就是农业机械化、优化播种以及精准施肥。

农业机械化

农业机械化的升级主要是指为传统农业机械装配卫星导航系统、传感器等先进的设备以采集海量的农业数据。气候数据有助于规划农作物的种植周期，科学选择农作物品种，土地遥测数据可以帮助规划土地利用比例，其他类型的数据则可以帮助进行其他的分析。

优化播种

智能化的农业机械可以自动地调节播种动作，以保证所有的种子都埋进相同的深度，具备相同的间隔，这样可以保持种子的利用效率，减少不必要的生长竞争与遮挡。

精准施肥

根据埋藏在田间的传感器，农民可以更好地了解庄稼的状况，根据其饥渴程

度，及时补水，补充必要元素。

5.7.2 制造

大数据助力制造业是 IT 人员视角下的一种讲法，而对于从事工业制造的人来说，他们更关注的是国家出台的“中国制造 2025”战略。但无论大数据制造还是“中国制造 2025”，区别并不明显。因此，对于制造业来说，可以分为以下四个方向。

故障诊断

制造企业中大型设备出现故障是难免的，而现如今的设备往往都已经具备了数字化控制机制，在运行的过程中会采集大量的设备状态数据，通过对设备数据进行分析并建模，可以做到故障检测与预警。

产品创新

在产品投入市场后，需要对产品的市场经营、销售、反馈等数据进行监控，以可视化分析的形式展现出来，从而实现数据驱动决策，产品情报分析，优化产品生产与资源配置。

工业物联网

能耗分析、质量事故分析也是在制造业中利用大数据进行辅助的应用。

供应链优化

在产品生产的过程中，需要涉及上下游产业链企业，有供应原材料的，也有下游销售服务体系。通过数字化企业管理，可以进行商品需求预测，提高仓储配送效能。

5.8 大数据行业成熟了吗

如果使用搜索引擎进行大数据市场报告的检索，你会发现很多企业以及各个地方政府都在发布类似的市场调研报告，这些报告看似由不同的组织发布，但其内容大体相似。除对国内外大数据整体状况进行概述外，最吸引人的便是过去一

年中各个行业的大数据发展情况以及投融资情况。

在阅读这些报告的时候，除关注各个行业的发展状况之外，还要找到每个行业的领军企业或者有代表性的企业，登录他们的网站、试用他们的产品、探索他们的商业模式，这是一种很有益的学习大数据在各个行业应用的高效行为。

本节中，我们将对大数据行业的成熟度加以讨论，另外还会介绍一些大数据的相关话题，对大数据从概念上进行收束。

5.8.1 行业成熟度

我们使用“应用成熟度”以及“市场吸引力”两个维度来衡量上面提到的行业。这里使用的衡量标准也是大多数行业报告中所使用的，应用成熟度可以从应用的普及程度上来判断，而市场吸引力则可以通过资本市场的投融资以及相关企业的营收状况来计算。粗略地，我们可以将一些常被谈及的行业在象限图中表示如下。

在图 5-5 中可以看出，各个行业的大数据应用成熟度可以分成三大阵营。

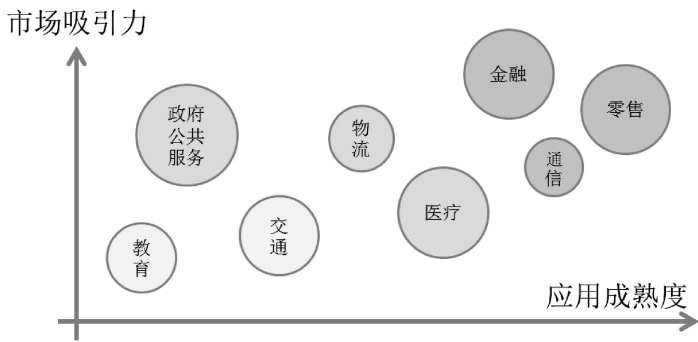


图 5-5 各行各业中大数据的成熟度

第一阵营是以金融、零售以及通信为核心的成熟阵营。特点是，有着成熟的商业模式，并且在产业链条上结构完整。从事该行业的投资相对来说收益可以预期。尽管存在风险，但是进入这样的行业会对未来有所期待。另外，一些技术被整合到这些行业后能够找到很多很好的应用场景。正是由于其是一个成熟市场，因此竞争也较为激烈。无论是在推荐系统、计算广告还是金融科技领域，都可以

列举出众多明星公司。

第二阵营是以政府服务、交通等为主的中等阵营。这个阵营中的行业往往以政府投资为主，属于基础设施投入。大数据技术在这些行业中的技术难度不高，但是想要形成成熟的商业模式较为困难。从产业链角度来看，这些行业往往会在某一个环节脱钩，尽管有 PPP 之类的模式可以借鉴，但是难免会有建而不用，用而不精的花瓶工程存在。加之数据的采集涉及隐私等话题，因而难以在这些领域建立数据的闭环，从而影响大数据在这些行业的应用效果。

第三阵营的代表行业是医疗、教育、制造业。在这些行业中，大数据的应用并不显著，主要还是以原始业务为主，推进的速度也相对有限。这些行业往往追求经验性，对于新兴的事物接受能力较弱。但正因为如此，第三阵营也是所有行业中潜力最大，投资回报率最高的。进入这样的新兴行业，尽管在大数据技术上有很好的积累，但是如果缺少了原行业的从业经验，也需要经过较长时间的摸索才能熟悉这个领域。在这个行业中最重要的是利用大数据技术作为辅助，服务原业务。

5.8.2 大数据理念

想必大家或多或少都看过关于大数据技术的介绍，那么一定绕不开大数据的 4V 特性。4V 指的是体量（Volume）大、速率（Velocity）快、种类（Variety）多、价值（Value）高。

体量大不用解释，现在笔记本电脑的内存都是几百 GB，甚至 1TB，而对于企业的业务来说，如果实施动态采集，数据量达到 PB 级完全没有问题。根据 EMC 公司的报告，到 2020 年，全球的数据将达到 44ZB。1ZB=1024EB，1EB=1024PB，1PB=1024TB。

对于速率来说，以前是单机离线处理，后来可以进行联机处理，再到后来是批处理，实时处理。现在对于 Hadoop 的大数据系统来说，可以通过一款叫作 Kafka/Flume 的软件实现流处理，大大加快了数据处理速度。

对于数据的种类来说，从之前的表格结构化数据，到后来的半结构化数据，再到包含音频、视频、图片等的非结构化数据，从单一的业务数据到网

络数据，再到现在的由“大物移云社”产生的各种物联数据，可以说真正实现了种类多。

数据的价值来则因业务而异，此处不做详细说明。

除了特性，如果要问大数据有什么样的作用，我想下面四条足以概括。

认知盈余

大数据从表象上来看是计算，有效扩展了人们的计算效率，使得人们有富余时间去思考更为深入的问题，提高了人们的认知盈余。从这个角度来说，大数据是一种高效率的计算工具。

逼近真相

大数据强调采集，由于采集数据的维度增多，原来采样下的世界由模糊变得清晰，犹如一块像素低下的图片被增强后一般，更逼近真相。

科学决策

大数据离不开数据挖掘，通过挖掘可以发现规律，而利用规律则可以指导科学决策。前提是你得相信大数据。

数据思维

大数据是一种方法论，是一种思维的升华，能够指导行动，赋予人们理性，这也就是我们常说的数据思维。

5.8.3 大数据趋势

在研究大数据技术的趋势时，我们时常会参考 Gartner 公司发布的技术趋势曲线。

在这里我简单摘取了 2013–2016 年共计四年的技术趋势曲线，让我们来看看这四年中大数据都发生了哪些变化。

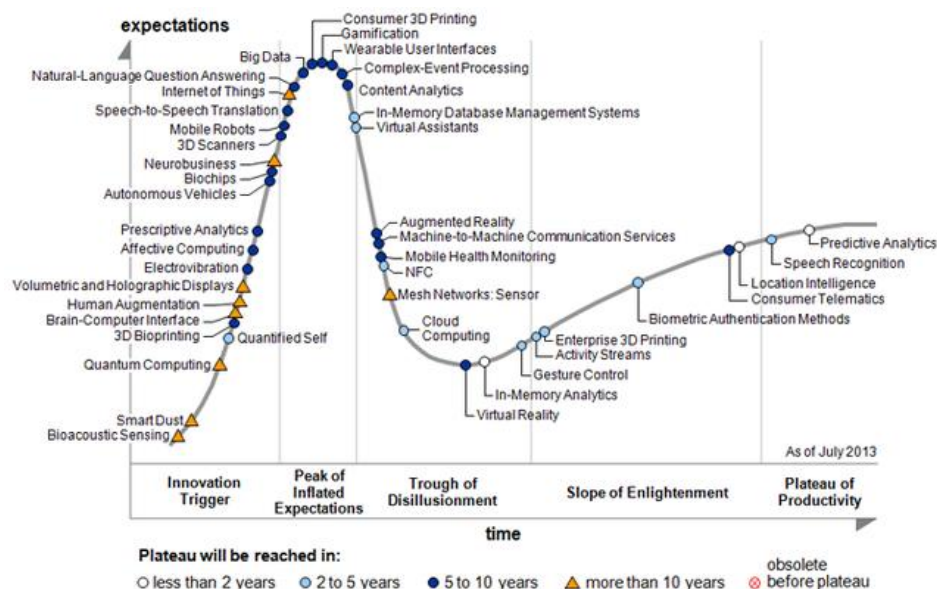


图 5-6 2013 年 Gartner 技术趋势曲线

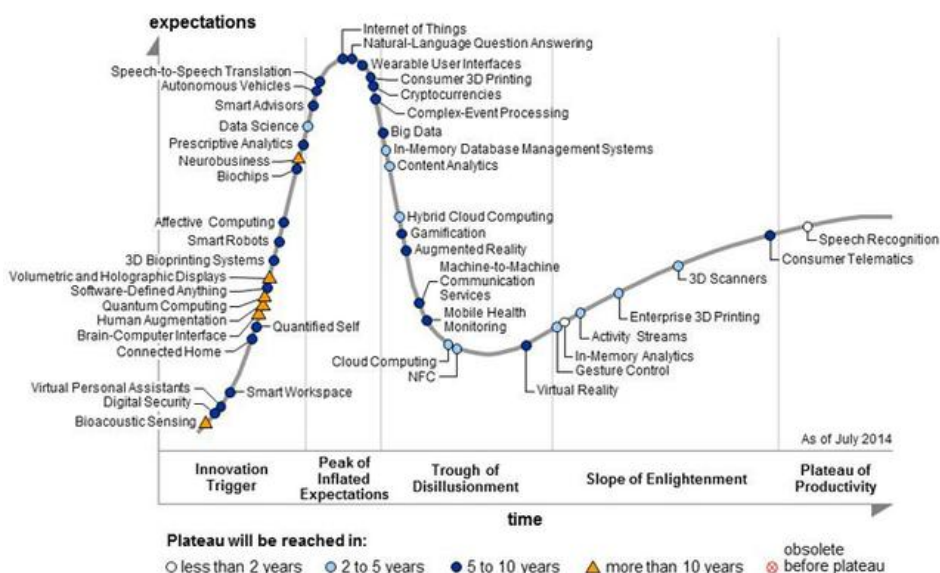


图 5-7 2014 年 Gartner 技术趋势曲线

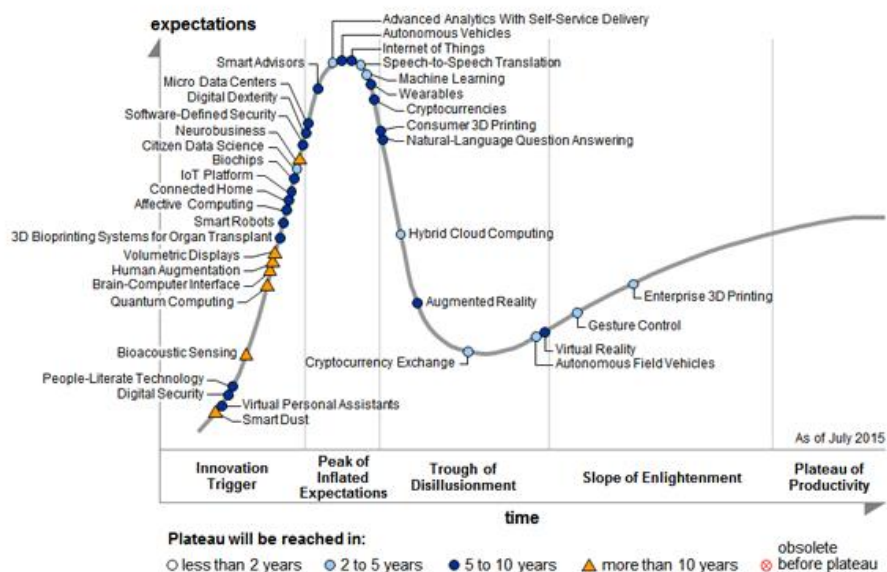
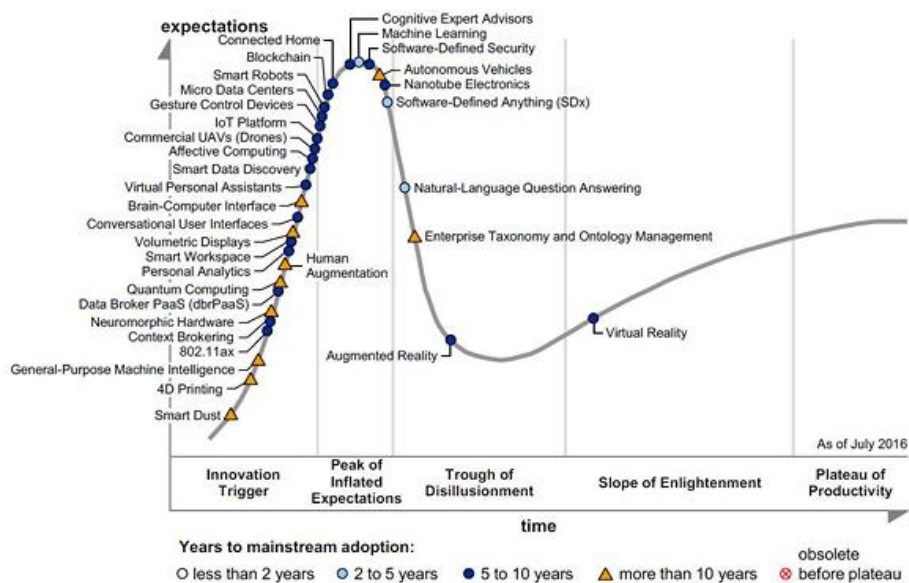


图 5-8 2015 年 Gartner 技术趋势曲线



Source: Gartner (July 2016)

图 5-9 2016 年 Gartner 技术趋势曲线

从上面四个年度的技术趋势曲线可以看出,大数据概念的热度在2013年到达顶峰,并在2014年进入了幻灭期,2015年与2016年间完全在曲线上消失了。而这样的消失并非是其影响力减弱,恰恰相反,由大数据培育出来的各个行业的应用正在“埋头苦干”,静悄悄地兴起。

从2015年与2016年两年的技术曲线来看,有三个趋势正在取代大数据浪潮。

一方面,以IoT为主的数据采集逐渐被重视,如2015年的IoT Platform(物联网平台)与Wearables(可穿戴设备),这是大数据的上游,它们采集到的数据正是喂养大数据的原材料。

第二方面,2016年的Machine Learning(机器学习)则是给出了大数据应用于各行各业的一个普适性解法,是大数据得以运转的原动力。大数据中的数据処理,以及与行业相结合的建模则离不开机器学习算法的支撑。

第三方面是由Blockchain(区块链)、Virtual Personal Assistant(虚拟个人助理)以及Personal Analytics(个人数据分析)所构成的,它们都处在很好的上升阶段。这算是大数据在各行各业中的落地应用,是大数据的下游。无论是数据采集的手段或者平台,还是各行各业的应用,都是围绕大数据这个概念发出的新芽。

5.9 大数据在产业中的位置

前面谈论了许多数据的思维、起源、概念以及在各行各业中的应用,接下来的论述将非常“烧脑”,不仅要求我们具有更为宏观的视角,还要有跳脱于技术的眼界。

作为数据产品经理,可以说是数据科学的圈内人与同行。不要妄自尊大,以谦虚的心态向整个人类社会学习,毕竟数据科学作为技术是需要依赖社会经济而存在的。也不可妄自菲薄,虽然数据科学是新兴学科,但是其历史由来悠久且脉络清晰,同时实实在在支撑着很多行业。有了这样的信念才可以说,我们已经准备好做一个数据科学的圈内人了。

下面从大数据行业组成与大数据在经济社会中的位置两个层面来介绍。

5.9.1 行业组成

FirstMark (<http://www.firstmarkcap.com/>) 是成立于 2008 年的投资公司, 总部位于纽约。每年, FirstMark 都会发布大数据行业从业企业的分布图(行业画像), 称为 Big Data Landscape。2017 年的大数据行业画像如图 5-10 所示。

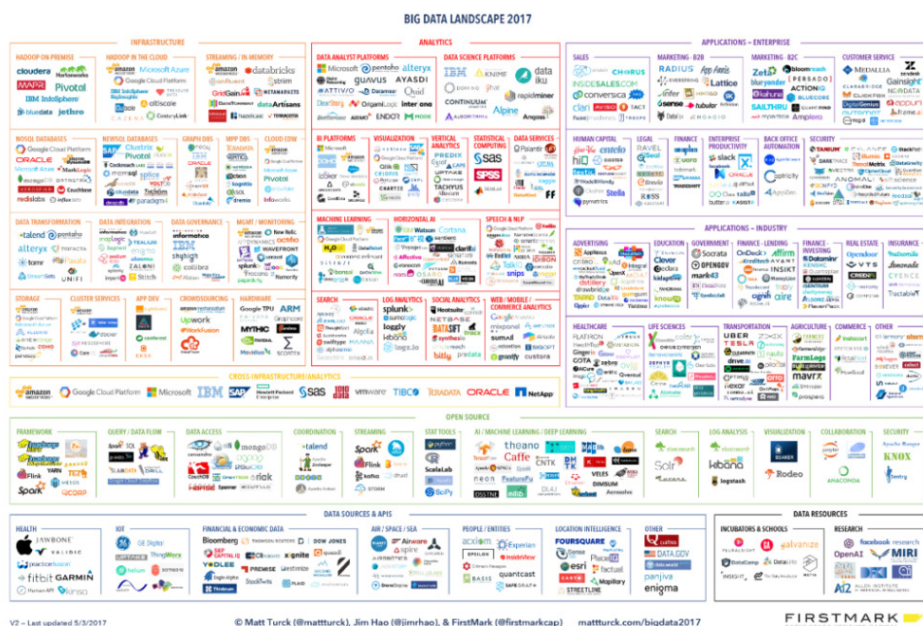


图 5-10 2017 年大数据行业画像

该图的高清版本可以通过 <http://mattturck.com/wp-content/uploads/2017/05/Matt-Turck-FirstMark-2017-Big-Data-Landscape.png> 获得, 除此之外图中提到的每家公司的数据也可以通过 <http://dfkoz.com/big-data-landscape/> 获得。

我们大致可以根据图中的各个模块来介绍大数据的行业组成。

数据源

位于底层的是数据源提供商或数据接口提供商, 可以说它们提供了供整个数据科学系统运行的原材料。这些原材料极大地丰富了整个大数据行业, 也为大数据行业大树的成长提供了厚实且有营养的土壤。

相关机构

与数据层相并列的，位于右下角的是 Incubator（孵化器）与 School（学校）模块。学校提供数据科学的技能培训，而孵化器则为整个社会的数据创业提供一系列的商业支持、资金支持以及技术对接指导。将孵化器与学校和数据源放在同一层可谓意味深重，它们共同构成了整个行业的基础。如果把数据源比喻成原材料和土壤，那么孵化器与学校就是辛勤的园丁和劳作的产业工人，是它们为整个生态物种提供了健康成长必不可少的氛围，让这个系统能够运转起来。

开源系统

在两者之上的是开源系统层。开源就是成熟技术模式的社会福利性公开或有偿性授权使用，这使得整个社会的接入成本降低。从图中可以看到，开源技术领域有专注于框架、检索、数据流服务、数据获取、实时处理、统计工具、机器学习、安全和可视化的各种组件，形成了一套完整的大数据解决方案。毫不夸张地说，开源系统是整个大数据之树的粗壮树干，支撑了所有的行业。

基础设施

在这株粗壮的树干之上，有三大领域，分别是基础设施模块，分析技术模块以及行业应用模块，如果不做严格的边界区分，可以认为是 IaaS、SaaS 以及 DaaS 的分别对应。基础设施是一切应用与分析启动的先决条件，在这个模块下，我们可以看到大家熟知的 Hadoop，Spark，数据库相关技术，数据迁移集成技术，安全与存储，App 开发与测试平台以及众包平台等。这些都是分析与应用的基础设施。

分析技术

对于分析来说，我们关注的是技术本身，包括分析平台型产品（初创公司、大公司均有）、可视化、商业智能、日志分析、社交分析、实时计算、机器学习、自然语言处理、泛人工智能、搜索引擎技术等。这些技术并不针对某一个行业或应用，可能会有更广泛的应用。

行业应用

区别于分析技术，应用模块更关注数据科学技术如何应用于各个行业。前文中已经细数了各个行业可能涉及的数据应用，但时间在发展，未来几十年的发展

可能远比过去几百年要快得多，因此要不断地更新自己的知识框架并向其中填充更多的内容。在图中我们看到，很多领域在之前已经涉及，如零售领域、广告营销领域、媒体领域、财经领域、教育领域、医疗领域、制造业等，也有一些领域尚未提及，如安全领域、垂直的人工智能应用等。

如果你对这个行业还抱有兴趣，那么为什么不自己跟进这个行业与领域的图谱呢？当新的报告发布时，不妨花上半个小时了解一下，然后找到自己所在的领域，找几家竞争厂商，将它们的产品与商业模式研究透，这样做对数据产品经理有百益而无一害。

5.9.2 产业构成

谈论完大数据行业，我们来看一下这个行业在社会产业中的位置。要知道产业是比行业更为宽泛的词汇。它包括政治、经济、科技与文化。举个例子，当我们谈论教育的时候，我们会想到什么？如果你是一个技术岗位出身的人，你的第一个念头或许是在线教育、教育信息化产品或者相关的技术应用，其次可能会是各个层面用户的需求。继续突破，看看这个领域除了技术还需要什么。教育领域还需要国家宏观调控的政策，在政治层面出台的相关政策，在教育经费上的投入以及各项事业的分摊，另外，你知道国家教育的支出占国家 GDP 的比重是多少吗？

根据世界银行（<http://data.worldbank.org/>）与国家统计局（<http://www.stats.gov.cn/>）公布的数据，2016 年各国 GDP 排名中，美国以 18.6 万亿美元位居第一，中国以 11.2 万亿美元居于第二位，其中，教育经费从 2013 年开始就已经突破了 3 万亿元人民币（约为 0.45 万亿美元），占国家 GDP 的 4% 左右，与国家公布的数字大致相同。

阿里研究院在这方面达到了业内新高度，已经近似于一家咨询公司了，其本身定位是国家智库。与之类似的还有腾讯研究院。图 5-11 是阿里研究院在 2016 年初发布的，这样的梳理在现在看来仍有借鉴意义。



图 5-11 DT 商业服务生态体系图

原图见 <http://www.aliresearch.com/blog/article/detail/id/20887.html>。

马云在 2016 年的阿里巴巴云栖大会上提出了“五新”，分别是新零售、新制造、新金融、新技术、新能源。这五个新都可以被囊括在新经济的框架体系内。而这五个新从内在逻辑上也是一环扣一环。新能源即数据能源，是基础，基于新能源需要新技术（如互联网、云计算、大数据与人工智能）进行支持，而新金融则需要更多的人参与技术创造，改造社会财富的分配方式，施行普惠金融，新金融必然导致新制造，而制造的产品需要进行销售，下一个阶段必然是新零售。当然，这五个新并无一定的先后顺序，有可能是同时起步，又相互耦合，螺旋上升，但无论如何，我们已经走在新经济的道路上。

再回过头来看阿里研究院的这张图，我们看到了从农业革命、工业革命起积累至今的农业、工业基础设施，也看到了数字革命以来建立的基础通信与互联网基础设施。随着数字革命的进一步发展，这些基础设施平台上涌现出一批新兴的用于嫁接其他科技成果的技术平台，如云计算、大数据等。可以说在新基础这个层面上，数据被理解成了基础设施。

基础搭建的目的必然是提供服务，没有提供服务的基础便失去了其存在的价值。首先是数据的开放，再向上是以此链接的生产、经营、金融以及物流等实实在在的用于指导新零售、新制造、新金融的服务，而这些信息化服务最原生的服

务又会被拆解与重组，拆解后变成了跨领域的综合服务，然后再次拼接并嫁接到各行各业，形成垂直服务。

对于新主体来说，原来市场中的消费者和企业的内涵不断扩大，囊括了各种小企业、自由经济体、网红以及独立制作人。在这样一个自由人自由联合的时代，新的主体必然激活新的经济。

所有的这一切都少不了实体规则与网络虚拟规则的指导，因此社会经济规则与网络规则看管着整个庞大与繁杂的新经济框架，使之平稳运行。

纵观整个产业，我们是如此渺小，却又如此伟大。渺小的是，我们仅仅像是一颗螺丝钉一样扎进了整个产业的机器之中，不停地运转，随时有可能被替换。而伟大的是，这个产业给了数据行业太多机会，我们只要努力一些就可以撬动更多，所以何惧之有？

第 6 章

当产品经理遇见数据思维

- 6.1 下一站：数据科学家 110
 - 6.1.1 数据科学的历史由来 110
 - 6.1.2 数据科学与商业智能 111
 - 6.1.3 数据科学的职业分类 112
 - 6.1.4 数据分析的技能进阶 114
- 6.2 数据产品经理的职业新要求..... 115

无论前面谈论的是处变不惊的数据思维，还是波澜壮阔的数据应用，最终都得回到真实的都市丛林，变身为“数据科学家”。

6.1 下一站：数据科学家

数据科学家是大数据时代最为热门的职业。对于从事数据科学的人来说，各个公司也给出了他们的要求，有三个定义令我印象深刻。

IBM 认为数据科学家是“一半分析师，一半艺术家”；埃森哲咨询公司认为“好奇心+分析能力+学习能力+业务+表现沟通+决策力”是从事数据科学这个行业的人员必备的素质；Facebook 则定义数据科学家的工作内容为“IT+统计+可视化+跨界”。

上述三家公司都提到了除硬技能之外的软实力，无论是“艺术家”，还是“决策力”，抑或“跨界”，这实际上都是对数据科学家提出的要求。这样的要求不是附属，而是和硬技能相互对等的，可见我们再也没有理由去排斥软技能。数据科学家的彼岸，不再是 CTO，而是 CDO、CIO，甚至 CEO。

接下来我们将简要地谈谈数据科学的历史，再介绍与数据科学相关的各个职业，最后聚焦到数据分析这个职位谈谈技能的需求。

6.1.1 数据科学的历史由来

数据科学一词最早出现在 1966 年，由 Peter Naur 提出，这位老先生也是 2005 年图灵奖（计算机界的诺贝尔）的得主。当时 Peter 提出这个概念的时候，数据科学不叫 Data Science，而是 Datalogy，充其量只能翻译为数据学，而不能称为数据科学。

那么数据学与数据科学之间的区别究竟是怎么样的呢？从某种程度上来说，数据学是研究数据本身，然而数据科学除了这个内涵之外，还肩负了为自然科学与社会科学提供数据研究新方法的责任。这说明在人类演化的过程中，数据的思维早已固化在大脑中，并被当成习惯，所以我们为了了解数据科学，也应该去了解自然科学与社会科学的发源，此处不再赘述。

6.1.2 数据科学与商业智能

对于任何一个学者来说，能够提出一个概念并被后人记住，远比当世的成就令其兴奋。从这个角度来说，Peter 老先生无疑是计算机界的赢家。伴随着数据科学概念的提出，后人围绕这个名词展开了广泛的讨论。大家都分别提出了自己对于这个概念的理解。

2005 年，美国国家科学委员会给数据科学下了一个定义，认为数据科学的作用就是“进行富有创造性的查询和分析”。其实，从 1966 年至今，有很多人尝试给数据科学下定义，但是之所以选取这个定义，一方面是因为美国国家科学委员会的地位权威，另一方面是其定义中提到了“富有创造性”这个词。数据科学不同于一般的 IT 工程，区别就在“创造性”这个点上。相较于其他项目中的问题已经被定义好，数据科学更需要自己找出问题。

在数据科学领域，被谈及较多的还有商业智能（Business Intelligence，BI）这个概念。下表从数据、分析对象、平台以及工具四个角度对两者进行了比较。

表 6-1 商业智能与数据科学特性对比

	商业智能	数据科学
数据	结构化	非结构化、半结构化
对象	群体性	个性化
平台	数据仓库、OLAP	Hadoop
工具	SQL、ETL	用户画像、推荐系统、其他算法等

数据

人们对 BI 的第一印象往往是各种仪表盘与数据报表，这个印象大致是正确的。BI 报表中的用户数、活跃数、留存数等情况，往往是对结构化的用户访问日志数据进行计数（Count）与求和（Sum）而得到的。而对于数据科学来说，印象中不应该停留在数据报表这个层面，许多很“酷炫”的应用都是数据科学的结果。例如微软亚洲研究院之前发布的人立方就是运用了人与人之间的社交关系做出的人的社交图谱，再如一些气泡图的可视化、用户标签的画像等都是数据科学的具体实践。因此，BI 更侧重于对结构化数据的统计，而数据科学则是对非结构化数据的分析。

对象

BI 往往是给企业经营者或市场运营人员来看的，关注的是全局。而数据科学更关注个体，更关注长尾，因此从推荐系统到客户细分都是在对用户进行个性化的分析与营销。

平台

若干年前的平台还是数据仓库（Data Warehouse）与关系型数据库的天下，基于数据仓库进行多维数据分析（OLAP）。随着数据量的增加，出现访问的负载均衡问题，主从分离、拆库、拆表等操作已经不能满足现有的需求，因而鲁棒性与可靠性更强的分布式平台 Hadoop 应运而生，关于 Hadoop 技术栈中的技术，可以参见专门的文章。

工具

以前从事 BI 相关工作的技术人员只需掌握 ETL 以及一些 SQL 语句即可，而现在的科学数据人员不仅仅要熟悉 SQL 语句（因为 Hadoop 平台中的 Hive 等技术也是用 SQL 语句进行查询的），而且还要对用户画像、推荐系统以及分类、聚类、回归算法有一些了解。

数据科学与 BI 没有谁将取代谁的趋势，更远的时期内两者将并存，并更好地融合在一起。从现在来看 BI 已经很好地吸收了数据科学的方法并为己所用。两者共生共存的一个主要原因就是，两者都是同一种数据思维，即使用数据驱动商业是一种闭环思维，本质上属于数据控制论的范畴。一言以蔽之（可能不太严谨），BI 用的是套路，数据科学需要的是创意。

6.1.3 数据科学的职业分类

生物里有“门纲目科属种”的划分，如果说数据科学是计算机科下面的一个属，那么隶属于这个属的种有哪些呢？大致可以分为两类，一类是分析，一类是实施。

在分析方面，有数据分析师、数据挖掘工程师、算法工程师这三个工种。从数据分析师到算法工程师，侧重点也从业务能力逐步过渡到了算法能力。

数据分析师

当数据越来越被各行各业重视的时候，数据分析师终于迎来了自己的春天，但是其面临的挑战也更加艰巨。相对于以前的数据匮乏，现在的境况是数据爆炸。这要求数据分析师在数据整理、分析、评估和预测等方面要下更大的功夫。其基本的分析技能包括对比、分组、结构、漏斗等分析方法，而高阶的技巧则包含了回归、聚类等数据挖掘技术。

数据挖掘工程师

如果说数据挖掘技术对于数据分析师来说是高阶技能，那么对于数据挖掘工程师来说，则是必备技能。回归、分类、聚类等算法自当烂熟于心，除此之外还要能够结合业务将这些技能用对地方。理解业务，抽象问题，建立模型，解决问题，评估问题，这是数据挖掘工程师日常工作的一个闭环流程。

算法工程师

数据挖掘工程师更多的是应用算法，而算法工程师有的时候需要改造甚至发明算法。不同应用领域的算法工程师拥有不同的技能。例如自然语言处理领域的算法工程师会了解分词、语义分析等工具，计算机视觉领域的算法工程师则会针对图像识别这一任务了解各种可以完成此项任务的算法。完成同一项任务的不同算法在算法工程师的眼中各有优劣，他们会根据实际情况来选择应用，并因地制宜地进行改造。

在实施方面，有运维工程师、开发工程师、架构工程师这三个工种。尽管这些工种在其他 IT 所属的领域中已然存在，但由于其与数据科学（特别是大数据）息息相关，相关技能需要迭代更新甚至全盘重来，因此我们在数据科学中再次提出。这些工种从运维到架构，技术能力逐步提升。

运维工程师

运维即运行维护，与开发工程师和架构师比较，他们以“酷”而取胜。他们每天打交道的是机房，是数据中心，需要做的事情就是保证服务器的运行效率，保证服务的可用性，确保成本控制与资源优化。网上打趣运维工程师的图片中最为典型的一张，要属众多运维工程师邀请大师在数据中心作法以确保服务稳定的画面了。



图 6-1 运维工程师请大师作法

开发工程师

这里的开发工程师是软件开发工程师的简称，包含前端工程师和后端工程师。前端工程师的工作成果用户直观可见，后端工程师的工作成果用户并非直接可感知。尽管运维和架构也需要从事编码工作，但是开发工程师是最常称自己为“码农”的人。

架构工程师

架构是一个抽象的词汇，在软件开发领域，架构师需要把框架搭建起来供大家填充内容。架构师需要选择软件开发采用什么技术方案，是以开源为主还是自主研发，是满足现有还是保证未来扩展，是轻量级还是大架构。简而言之，架构师是登峰造极的码农，也是通往 CTO 的路径之一。

6.1.4 数据分析的技能进阶

了解数据分析的工作内容，对于数据产品经理来说十分必要。从某种程度上说，这就应该是数据产品经理的职责内容之一。

对于数据分析师来说，我认为有三个要求需要做到。分别是言之有据、言之凿凿以及言之有物。言之有据是指使用量化数据替代平时的“拍脑袋”，这是数据思维进化的第一步。言之凿凿是指使用真实的数据说话，从用数据说话

到用真实的数据说话，体现的是对数据的较真精神。不仅蕴含了严谨的态度，也使得业务的输出有保障。言之有物是更高层次的要求，不仅要在数据层面得出结果，还要能够根据表象给出一些意见和建议，可谓是使用真实的数据说有意义的事。

我们还可以从另一个角度来看数据分析师的技能，可以概括为“是什么”“为什么”“会怎样”与“怎么办”。对于“是什么”，涉及的技能有数据报表（Excel）、即席查询（SQL）、多维钻取（透视表和 OLAP）、数据预警。当问题进入“为什么”的时候，就不是共性的工具可以解决的了，这个时候需要调用统计学知识进行数据分析。“会怎样”是基于历史数据进行预测的一个过程，统计学知识需要升级为数据挖掘的技巧，如回归模型。而最高级的形式就是“怎么办”，这有点类似于上述的软技能，也可以和言之有物相对应，需要的是逻辑思维能力。

当我们趟过人类社会与 IT 漫长的历史河流，职业规划就是崎岖不平的河床，在脑海中逐渐清晰。选择一条支流，专注于数据，在这条涓涓细流的沿岸深耕细作，相信这片肥沃的土壤必将培育出丰硕的果实。

6.2 数据产品经理的职业新要求

面对汹涌的数据大潮，产品经理们自然是应该惶恐。首先，没有人知道他们应该做什么。按照传统的观点来做自然不会出错，但是显然也做不出特色。如果摸着石头过河，又会因为试错而损耗一些激情。其次，没有人告诉他们做得对不对。假设产品经理们面向公司的大数据战略做出了改变，并且改变的方向是附和行业与产业方向的，可是谁能保证有伯乐会识别这匹好马呢？最后，假设上面两个问题都不存在，你知道该做什么且有人支持你做，那么做成之后的收益是什么呢？对于团队成员来说，这不过是一个再普通不过的项目，其价值恐怕无法被广泛认同。

伴随问题而生，一定有解决方案。我在公司内做的数据产品经理培训便是在寻求这样的方案。我们细数了数据产品经理可能涉及的各个领域，以及在每个领域应该了解的深度，目的有两个，一个是促进工作更加高效地进行，另一个是给

每个人更多的技能上升渠道。

在大众看来，数据产品经理就是产品经理与数据分析师的合体。其实，除了两者的拼接，可能还需要一些粘合剂才行。我在各个公司的主页以及各大招聘网站上查阅了与产品经理及数据分析师相关的工作岗位需求。我们调研的公司包含了当今主流的 BAT（百度、阿里巴巴、腾讯）以及未来的潜力股 TMD（今日头条、美团、滴滴）。让我们来看看各公司是怎么要求这个职位的。

表 6-2 百度数据产品经理招聘

工作职责	<ol style="list-style-type: none">1. 负责百度数据服务类产品的规划、需求分析和产品设计2. 关注大数据应用相关方向的前沿研究，并将成果快速产品化、商品化，将创新推向亿万用户3. 负责项目推进过程中的跨部门协调沟通工作，能够协调各资源以确保产品顺利发布
任职要求	<ol style="list-style-type: none">1. 硕士及以上学历，3 年以上相关工作经验，具有计算数学、数理统计和计算机相关专业背景者优先2. 有比较丰富的产品经验，在需求的确认、产品的原型，PRD 的撰写、资源的协调，产品上线后的跟进与项目管理等方面有实践经验3. 具有数据挖掘分析功底和敏锐的数据洞察力4. 信仰、热爱数据

表 6-3 百度数据分析师招聘

工作职责	<ol style="list-style-type: none">1. 负责业务相关数据整理及分析工作，提交日常数据分析报告2. 通过对数据的深度分析，规划、开发运营分析指标体系，为运营决策等方面提供数据支持3. 通过对数据的敏锐洞察，发掘潜伏的异常现象并迅速定位问题本质，提出解决方案4. 根据数据分析模型，生成数据分析报告
任职要求	<ol style="list-style-type: none">1. 两年以上数据分析、挖掘和建模经验，数据分析相关方向2. 具有优秀的数据分析能力，能快速理解业务并设计相应的数据模型3. 熟练使用 office 办公软件，至少熟悉一种数据分析软件（SAS、R、Clementine），独立完成数据分析报告4. 有互联网相关行业及产品运营分析工作经验者优先

表 6-4 阿里巴巴数据产品经理招聘

工作职责	<div>1. 深入理解和把握内部外部业务需求并进行业务规划，持续对产品进行优化迭代，监控从规划设计到测试发布的全过程</div> <div>2. 负责公共出行大数据分析产品的产品规划、管理工作</div> <div>3. 负责大数据产品的调研、分析及产品设计</div> <div>4. 与资源团队紧密配合，快速、高效推动产品设计、研发、运营、推广</div>
任职要求	<div>1. 具有移动客户端或 Web 端产品设计及管理经验，熟悉移动互联网产品和服务</div> <div>2. 具备优秀的需求分析和产品规划能力，以及独立的业务分析、数据分析、竞争分析能力和见解</div> <div>3. 沟通、执行能力强，乐意接受有挑战性的工作</div> <div>4. 具有追求卓越的耐心及热情，乐观积极</div> <div>5. 及时掌握行业最新动态，关注行业、业务发展趋势，关注创新性产品理念，并对关键点有独立见解</div> <div>6. 有交通领域背景或大数据产品经验者优先</div>

表 6-5 阿里巴巴数据分析师招聘

工作职责	<div>1. 针对互联网行业进行长期追踪分析，在一些创新方向上有敏感性，如 IoT、AR 等</div> <div>2. 可以提出有价值的分析模型，配合挖掘/算法工程师的输出结果，针对项目进行挖掘性分析</div> <div>3. 能独立完成相关行业数据的收集、提取，规划并完成相关行业报告</div> <div>4. 了解行业客户的痛点，提供有价值的分析结论，并尝试商业化</div>
任职要求	<div>1. 熟练使用数据统计分析工具，能快速有效地给出数据分析方法并得出结论</div> <div>2. 熟悉 SQL 语言，对大数据进行筛选过滤，快速处理数据；有 Hive、ODPS 等平台使用经验者优先</div> <div>3. 思维敏捷、对大数据具有敏锐的洞察和分析能力，有创新意识，能独立编写专业的行业数据分析报告</div> <div>4. 了解互联网趋势，熟悉大数据发展方向及市场现状</div>

表 6-6 腾讯数据产品经理招聘

工作职责	<ol style="list-style-type: none"> 1. 负责腾讯互联网大数据产品的运营和策划 2. 制定产品中长期目标，盈利模式和推广策略 3. 负责产品版本规划，原型设计，组织需求评审，推进产品敏捷迭代和上线工作 4. 挖掘用户需求，用户体验改善，操作流程优化，建立产品和数据质量的验收标准，持续优化和打磨产品 5. 跨部门/公司内外部合作沟通，追求产品价值最大化
任职要求	<ol style="list-style-type: none"> 1. 5 年以上互联网公司工作经验，3 年以上产品运营/策划工作经验 2. 熟悉移动互联网行业数据产品，对行业动态变化敏感度高，执行能力强，有较深厚的产品文档撰写能力和交互能力，以及产品包装能力 3. 较强的团队协作和沟通能力，思维活跃，学习能力强，适应能力强 4. 有强烈的价值驱动业务的使命感，工作主动积极，能承受一定工作压力 5. 具备移动 App 数据分析、市场和用户行为研究、投研分析等经验者优先

表 6-7 腾讯数据分析师招聘

工作职责	<ol style="list-style-type: none"> 1. 根据公司业务设计有效的数据指标体系，进行运营数据跟踪分析，发现潜在的商业机会、风险或缺陷，提供决策数据支持 2. 对海量商业数据进行挖掘与建模，建立有效的客户、产品分析模型供业务部门数据化运营使用，提高效率，扩大收入 3. 负责各个垂直领域的业务数据分析，垂直领域画像等相关底层数据的建设，协助构建完整的用户画像体系 4. 负责提供数据模型设计和代码文档，参与数据清洗和过滤、数据处理、数据分析可视化等工作
任职要求	<ol style="list-style-type: none"> 1. 具备熟练撰写内容丰富的可适用于公开演讲级别的报告的能力 2. 具有优秀的沟通协调能力及开拓精神 3. 具有良好的撰写分析报告的能力，熟悉常见的统计原理及方法 4. 熟悉统计学基础知识，掌握数据分析的体系流程与方法 5. 具备较强的数据敏感度，善于收集整理行业数据，熟悉及关注互联网行业的发展动态 6. 熟悉 Hadoop、Spark 等分布式计算和存储平台，会编写 Hive SQL、MR 分布式计算程序 7. 熟悉 Shell、Python、R 等语言及相关库包，熟悉 Linux/Unix 平台的开发环境 8. 具备缜密的逻辑思维能力，敏锐的观察能力和独立分析能力，有较强的市场感知力和数据敏感度

表 6-8 今日头条数据产品经理招聘

工作职责	<ol style="list-style-type: none"> 1. 参与公司核心业务策略设计，开展多维度业务分析工作 2. 通过数据挖掘进行客户画像、市场大盘分析、产品资源分析等工作，寻找改进点与创新点，提出业务策略建议 4. 参与广告产品分析工作，从不同视角发掘产品潜力，推动产品改进 5. 在日常工作中，处理反馈各类相关需求
任职要求	<ol style="list-style-type: none"> 1. 统计学、应用数学、计算机相关专业本科及以上学历，研究生及以上学历者优先 2. 两年以上互联网相关经验，熟悉销售策略、广告策略者优先 3. 有数据分析及建模相关经验，如客户画像、战略规划、商业变现分析等 4. 熟练使用 Python、Axure、SQL、Excel、PPT、SPSS 等工具者优先，能够独立编写数据分析报告并完成产品设计 5. 具备良好的沟通交流能力，较好的逻辑分析能力，敏锐的商业嗅觉，对数字敏感

表 6-9 今日头条数据分析师招聘

工作职责	<ol style="list-style-type: none"> 1. 数据管理：梳理业务逻辑模型，建设主题表，埋点验证等 2. 报表支持：管理和建设报表体系 3. 业务分析：根据用户行为数据进行业务分析，给业务端更好的指导，帮助业务快速提升，同时建设业务分析模型，不断评估和优化业务分析模型
任职要求	<ol style="list-style-type: none"> 1. 具有数学、统计学、计算机相关专业背景者优先，对数据分析有强烈兴趣者优先 2. 有 IT 大数据分析经验，具有大型互联网公司数据分析经验、互联网数据建模分析经验者优先 3. 有一线产品或运营业务优化经验者优先 4. 精通 SQL，熟练使用 R/Python 进行数据清洗、分析 5. 能快速理解业务，发掘业务细节和数据之间的联系 6. 掌握 Tableau 等 BI 工具尤佳

表 6-10 美团数据产品经理招聘

工作职责	<ol style="list-style-type: none"> 1. 理解公司战略和业务发展方向，梳理各业务部门数据需求并形成有效的数据产品方案，对数据产品有 owner 意识 2. 规划配送平台数据产品，制定阶段性产品目标
------	---

续表

工作职责	<ul style="list-style-type: none"> 3. 跨部门协调沟通，有效推进项目进展，完成既定产品目标 4. 监控现有数据并对异常数据进行分析，协调研发部门处理异常情况，梳理总结形成流程或规范以便前瞻性解决类似问题 5. 熟悉现有数据产品，针对新上线的数据产品组织培训，协助业务人员理解和运用
职责要求	<ul style="list-style-type: none"> 1. 5 年以上工作经验，统招本科以上学历 2. 3 年以上互联网数据产品相关工作经验，有快递、配送或 O2O 行业数据背景者优先 3. 条理清晰，责任心强 4. 具有良好的跨部门沟通协调能力，学习能力强，工作积极，能承受较大工作压力 5. 熟悉 SQL 查询，熟悉 Hive 操作，了解 Axure

表 6-11 美团数据分析师招聘

工作职责	<ul style="list-style-type: none"> 1. 负责业务研究项目。将业务方的需求经过数据分析落地为可执行的策略或方案，并追踪执行结果 2. 搭建主题性的指标体系，并负责数据报告的制作 3. 培训运营人员，提高他们的数据分析能力 4. 应用并优化现有数据工具，同时寻找新的工具以提升生产力
职责要求	<ul style="list-style-type: none"> 1. 具备“需求分析→数据提取和清洗→分析建模→数据可视化→报告撰写→决策支持”全流程的分析工作经验 2. 具备优秀的商业分析思维，善于思辨，敢于决断。能够针对市场营销业务搭建整个数据平台 3. 熟悉市场营销业务，可根据数据分析和推演进行预算的拆解和营销策略的监督执行，擅长分析总结，将策略和经验前置置于业务 4. 具有数据提取及处理能力。能通过 SQL、Excel、R、Python 等工具提取并处理原始数据，使其贴合分析需求 5. 具有数据可视化能力。能通过数据分析工具制作精准有效的数据图表，传达准确的信息 6. 具有信息挖掘能力。熟悉数据分析的方法和业务，能通过数据透视、数学建模和逻辑推理从数据中解读出有价值的信息 7. 具有熟练的商业报告撰写能力 8. 具有强大的自驱力，具备分享精神，能带动小伙伴们一起提高

表 6-12 滴滴数据产品经理招聘

工作职责	<ol style="list-style-type: none"> 1. 负责与平台开发团队一起规划和完善数据平台，完成需求的收集、沟通、规划、项目推进、效果验收、线上故障跟进、用户反馈与产品运营等全过程 2. 负责产品设计，包括数据平台各子系统（报表、Ad Hoc Query、OLAP、数据可视化等）的功能设计 3. 承担平台、工具、系统相关的运营工作，与用户建立联系，收集和梳理用户反馈，及时调整产品策略，及时进行功能升级 4. 与分析师共同完成基础维度指标体系的搭建（数据需求的分析，指标、维度定义）
职责要求	<ol style="list-style-type: none"> 1. 具有计算机、统计学、数学等相关专业背景，本科及以上学历，两年以上数据产品规划和设计工作经验 2. 熟悉互联网产品整体实现过程（从需求分析到产品发布），深刻理解用户需求 3. 熟悉 BI、DW 原理和实施，有 BI 项目经历，有数据仓库构建和使用经验者优先 4. 责任心强，对数据的准确性、及时性及稳定性有认知和要求，有数据治理、数据质量监控经验尤佳 5. 有较强的逻辑思维能力，良好的沟通协调、执行推动能力和数据分析能力，对数据敏感 6. 掌握 SQL，能独立完成基础的数据探查工作 7. 熟练掌握 Axure、Visio、MindManager、MS Office 等基本工具

表 6-13 滴滴数据分析师招聘

工作职责	<ol style="list-style-type: none"> 1. 基于业务逻辑建立业务数据模型，为团队设计反映产品和运营状况的数据报表 2. 持续跟踪业务数据，监测业务发展态势，为业务指标异常提供科学解读 3. 参与滴滴数据体系建设工作，包括但不限于数据埋点流程设计、数据产品内容建设等 4. 结合行业数据、业界观点和市场趋势，为长期业务发展方向提供战略层面的意见和建议
职责要求	<ol style="list-style-type: none"> 1. 对数据敏感，逻辑严谨，责任心强，工作主动自驱，沟通能力强 2. 有一定的互联网行业工作经验，对互联网产品和服务兴趣浓厚 3. 受过基本的统计学训练，有一定的数据库基础，熟练使用 SQL 4. 熟练使用 Excel、SPSS 或 R 等数据分析和统计分析工具 5. 有建模、ABtest 经验者优先

表格罗列了上述几家公司对数据产品经理与数据分析师的工作职责及能力要求，其中共性的能力关键词进行了突出显示。将众多技能集中在一张图上，如下所示。

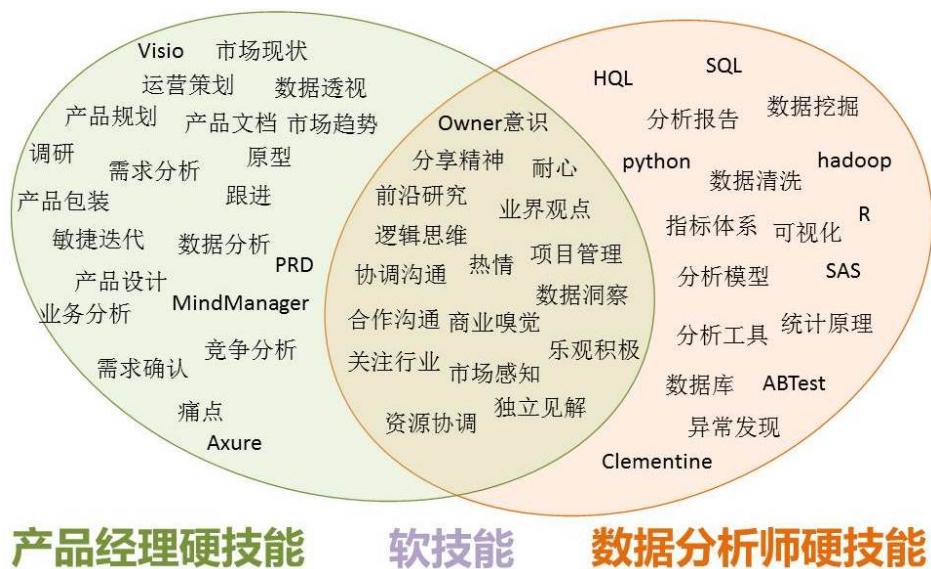


图 6-2 产品经理与数据分析师关键技能

从上图可以看出，数据产品经理的技能要求很清晰，大致可以分为三个部分，分别是产品经理硬技能，数据分析师硬技能，以及作为产品经理与数据分析师都要具备的软技能。

对于硬技能而言，我们看到了熟悉的产品设计、需求分析、原型设计、竞争分析等技能点，可以说这是产品经理的安身立命之本。

对于数据分析师来说，除了需要了解一些程序语言，还需要了解诸如指标体系、可视化工具、分析工具等知识，可以说是半个 IT 人员。

而中间的软技能包括市场、组织能力等。

很多时候，我会问自己一个问题，数据产品经理到底是具备数据分析能力的产品经理还是拥有产品思维的数据分析师？事实是，任何人都可以转变为数据产品经理，只要具备上述能力，而这些能力又因人而异，可以进行调整，很多非相关专业出身的人也可以胜任，毕竟产品经理是一个岗位，而并不是一个专业。

但有一点是可以确定的，数据产品经理一定是一个通才，而不再是只会画原型图、写 PPT 的专才了。这个社会已经是一个跨界的社会，很多财富藏在了金字塔的中间地带，如果不选择跨界，不选择用好奇心和求知欲去了解，那么就注定只能在金字塔上爬行，而无法前往金字塔中间地带。回到数据产品经理的主题上，难道你不觉得这个岗位提供了跨界的机会吗？

最后附上图 6-2 所示的数据产品经理技能图谱。以产品基本技能为底，上面的支撑包含统计、数据挖掘、可视化、平台以及运营等其他技能，再辅以思维、文案和其他软实力做顶，构成了完整的数据产品经理的必备技能图谱。

接下来，数据科学领域的标准技能即将拉开帷幕。

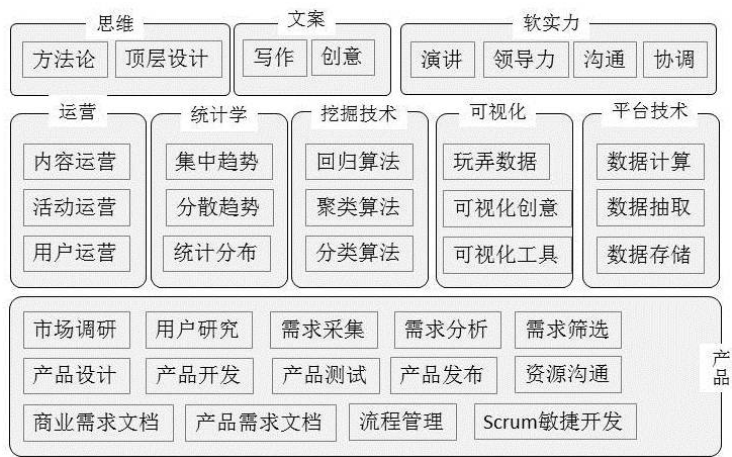


图 6-3 数据产品经理技能图谱

part three

第三部分

每个门派都有自己的武功和绝学，每项武功都有属于自己的一部“四库全书”或是“吕氏春秀”。如果非要在数据分析人员的武林秘籍上写上必备能力的话，我想至少是数据预处理、统计分析、数据挖掘与可视化。

“自古真情留不住，从来套路得人心”，数据分析也有一个标准姿势。但遵循套路不是目的，客户满意与效益达成才是评判好坏的标准。

这个部分将介绍数据分析和数据挖掘，除此之外，和数据相关的运营、研发与销售也将提到。

数据分析与挖掘本就与统计学紧密相连，展示若干数学公式也无可厚非，但我并不打算用数学来表达，而是期望用加减乘除与历史故事来讲清楚。让我们轻松上路吧。

数据产品经理的技能进阶

第7章

面向产品经理的数据预处理

- 7.1 数据分析的标准姿势 128
- 7.2 淘洗数据沙砾（数据清洗） 130
 - 7.2.1 缺失值 130
 - 7.2.2 异常值 132
 - 7.2.3 归一化 133
- 7.3 聚细沙成佛塔（数据集成） 135
 - 7.3.1 实体识别 135
 - 7.3.2 冗余性识别 136
- 7.4 换个姿势再来一次（数据变换） 137
 - 7.4.1 离散化 137
 - 7.4.2 属性构造 139
- 7.5 少即是美（数据规约） 139
 - 7.5.1 特征规约 140
 - 7.5.2 样本规约 141

数据分析和挖掘领域有一句话叫作“Rubbish In, Rubbish Out”，意思是，对于算法和模型来说，如果输入的数据是低质量的，输出的结果质量也高不到哪里去。为此，在开始一切工作之前，需要对数据进行预先处理。

千里之行，始于足下。让我们先从数据预处理开始。数据预处理中最传神的字就是“预”。在我看来，大致有三层含义。一方面，预表示预先，指明了处理的时机，即在一切事情之前做的。事实上，在建模和评估之前进行数据预处理就是在进行 todo 阶段的准备工作。另一方面，预还有预备的意思，即从技术上为接下来要进行的工作做一些准备，这样当项目开展起来的时候，不至于手忙脚乱。在这个意义上来说，这个过程越充分后面的过程就越顺利，正所谓良好的开端是成功的一半。第三方面，预还有预热的含义，即整个团队在思想上经过了 kick-off（项目开机仪式）之后能够重视起来。

预处理的内容也包罗万象，大致说来包括清洗、集成、变换与规约。很多书籍中会将一些具体的问题分别放入这四个预处理的环节中，尽管对于一些具体问题，可能放入的环节有差异与可探讨的空间，但如果你不是一个理论派的话，相信这些区别不会影响你学习新的技能。

7.1 数据分析的标准姿势

说实话，数据分析根本没有什么标准姿势。而为什么网络上有各种数据分析方法呢？那是工业社会机械化生产的产物。不妨让我们这样思考：假设你是人类社会第一个进行数据分析的人，根本谈不上方法，很多情况下你甚至没有意识到自己在进行数据分析就做了一件开天辟地的事情。当你遇到越来越多需要进行数据分析的场景，你就会发现它们之间的共性，你大脑中的机制这个时候开始启动，这些机制使得自身的能量消耗降低，而其具体表现就是把类似的经验在大脑中固化下来形成规则，也就是网上看到的方法论以及各种步骤。所以方法实际上不过是人身体生化运转过程中的一种工业化进程罢了，它能够节省能量，使得大脑有更多的余量去思考别的问题。很多时候我们之所以没有自己的方法，不过是遇到的实际场景太少，或者是大脑没有能量富余而使得总结升华为规则的效率降低而已。

这里给大家介绍一种新的数据分析框架。严格意义上来说，这个框架是数据挖掘的框架，可是数据挖掘和数据分析在业务层面上并不分家，因而可以这样迁移。具体来说，其过程如下。

商业理解

主要是理解业务需求，这个需求可以来自于客户也可以来自于上级，但有什么区别呢？某种程度上来说，上级就是你的客户，你需要像管理客户需求那样管理他们的需求。在这个阶段，要明确目标，并且站在客户的角度换位思考。需求虽从客户嘴中说出，但是“文不对题”的现象严重，需要透过现象看本质，了解其动机，进而完成商业理解。这种理解我们会在产品技能部分再做详细介绍。

数据理解

这个过程主要是进行数据的收集与初探。收集是了解数据放在哪里，而初探是对数据进行大致的了解，如记录条数、时间跨度、维度属性等可以在这个过程中了解。在这个过程中，可以使用两人一组的方式进行数据理解，我称之为“苏格拉底之问”。之所以称为苏格拉底之问，是因为苏格拉底提出了质疑一切的哲学思想观念，通过由一个人准备数据，另一个人不停地对准备数据的人进行口头询问，促使准备数据的人了解自己准备的不全面之处，也可以两人同时准备，相互发问，几轮下来，数据的底便可以摸清楚。

数据准备

这个过程中数据需要被清洗，如去重、去异常、规约等，详细内容会在后续章节中介绍。

建模

建模就是建立模型，在这里，我们优先指的是业务逻辑模型，但业务逻辑模型并不落地，所以其背后可能站的是算法模型。这里不涉及数据库模型。

评估

对模型进行评估，这个过程是算法模型过程中的必要一步，也就是为了证明给别人看，你的模型是优秀的。这个评估阶段可以是自证阶段，即只要自圆其说模型好用即可。

部署

部署便是真实落地，其内容包括可视化、工程化或报告呈现。没有这一步，前面做再多只有苦劳而没有功劳，只有落地且形成销售的工作才是优秀的工作。这是以终为始的必然要求，也是目标为导向的职业素养。

上述六个过程有一个名字，称为 CRISP-DM，即 cross industry standard process for data mining，其本质离不开我们的 to do、do、done 的框架，从这个意义上来说，商业理解、数据理解、数据准备是 to do，建模与评估是 do，而部署则是 done。

大家不妨翻开自己曾经收藏的关于数据分析流程、方法、框架的文章，所有的内容无非就是这些，但这些方法在解决实际的问题时却怎么也没有办法 100%地套用进去。探索的路就是这样，总是在不断更正航线中到达彼岸，而这个时候你也就从一名水手变成了船长，从一名乘客变为了司机。

7.2 淘洗数据沙砾（数据清洗）

对于进行数据价值探索的产品经理来说，大部分数据就像砂砾一样，要从中找到金粒一般有价值的数据，才能够做出卓越的数据产品，这也是“数据挖掘”一词的来历吧。

在海量数据中，大部分的数据是有缺陷的。这样的缺陷主要体现在数据项有缺失值，数据值出现异常，数据之间没有归一化从而不具有可比性等。

7.2.1 缺失值

在很多产品中，用户信息中心是建立用户体系必不可少的页面之一。在这个页面上，用户被询问芳龄几何、姓甚名谁、收入范围、性别归属等问题。这些数据对于产品来说往往用于用户结构分析，因而并非必填项目。如果填与不填获得的产品与服务前后差异不大，用户往往会忽略填写，甚至进行一些恶作剧，产生异常值，造成数据缺失。

对待数据缺失，要分三步走。分别是定、删、补。

定范围

定就是要定性与定量了解数据的情况。数据是对客观世界的描述，而客观世界又是存在约束的。

对于已经收集回来并存储在数据库中的数据，了解数据库中哪些字段有缺失，缺失比例如何，这是一种定量的了解。明确有缺失数据的字段重要性如何，这是一种定性的描述。定量的描述相对容易进行，只需要寻遍所有数据库中所有记录的属性即可。定性的描述则需要与业务场景相结合。

与此同时，还需要根据上述问题确定策略，是填充还是使用其他渠道进行补齐？是直接去除还是忽略？这就是我们接下来要用到的两个策略。

删字段

缺失的数据项就像是后进的士兵，要么丢弃，要么拉他一把。删字段便是丢弃的策略。很多情况下，我们总是希望数据越多越好，删除数据有悖常识，因而删数据也是需要一定魄力和勇气的。

我们的底气来源于哪里？主要有两个方面，一是对业务清晰的判断，二是对数据“有心杀贼，无力回天”的感叹。第一种情况下，如果一个字段对于后续的业务没有太多的帮助，便可以直接删除。对于第二种情况，即便这个数据项目对业务很有帮助但是难以直接或通过间接方式补齐，也只能作罢。对于这种情况，仅有极少量样本在该字段上有数据，则可以采取少数样本特例讨论的方式抓取出来单独进行研究，而对于整体则选择直接删除这个字段。

补数据

对于数据补充来说，有三个方式：一是业务知识/经验填充。如果我们需要分析的是 K12 的教育数据，并且知道学生的年级，那么大致可以根据学生 6 至 7 岁上学这一业务知识对相应年级学生的年龄缺失情况进行补全。二是使用均值、中位数、众数进行填充和补全。以均值为例，我们可以使用全体用户的收入均值来补全那些尚未填写收入的数据。三是使用其他渠道补充。很多数据包含一些隐性的意义，例如手机号可以反映用户的归属地。再例如，身份证号中包含了年龄信息等。

任何场景都是讲究特殊情况的，特殊情况特殊分析，随机应变才是贯穿数据处理的主线。

7.2.2 异常值

收集到的数据来自于采集，如果是系统自动采集还好，但如果是用户自己填写，就很难保证所有的数据都是统一的格式了。譬如让用户输入时间，不同的用户可能会有不同的理解，有的用户给出的是年月日顺序的 8 位数字组成的字符串（如 20170907），欧美输入习惯是日月年（如 7/Sep./2017）。不同的输入会给数据处理带来不同问题。很多系统提供统一的日期选择组件来解决用户输入层面的不统一。但是当不同的系统进行数据融合的时候，不同的日期存储形式还是会产生冲突，一种解决的方法可以是转换成统一的时间形式，如 UNIX 时间戳（计算从 1970 年 1 月 1 日 0 时起到现在为止的秒数）。

上面描述的异常只是其中一个例子。从整体上来说，数据出现异常主要体现在四个方面，分别是格式、字符、合理性与一致性。下面从这四个方面分别进行阐述。

格式

对于格式来说，有时间日期、数值以及其他一些固定格式的情况。例如对于时间和时期来说，可能会出现“2017-3-14”“2017/3/14”以及“14/Mar/2017”等多种情况，遇到这样的情况就要使用统一的方式来进行规整，使之成为统一的格式，如上述提到的 UNIX 时间戳。还有一些因为人为输入习惯造成的数据差异，如“20”与“二十”，甚至在一些情况下有可能被表示成“0020”，如果不在字段中找到这些区别并将其进行格式上的规整，后期处理时将会被作为不同的数值对待，影响结果。

字符

对于姓名来说，有的时候可能有半角空格与全角空格的区别，例如将“张三”写成“张 三”，中间多了一个空格，使得计算机难以判断两者是同一个人。很多时候由于用户填写错误或者后台处理程序读取与编写的错误，使得姓名、年龄、手机号码等位置错乱，如把年龄填写到姓名字段中去，这也势必使数据分析多了一层障碍。

合理性

很多时候数据并没有缺失，也没有填写错位，但是却由于不符合常理而被审

查清洗。例如，年龄填入 200，再如年级是 2020，或者手机号码是 13000000000（没有意义）。审查数据的合理性需要结合业务进行。一般情况下，产品经理可以对数据进行一些常识性的推理，但在具体的业务场景下，还应该尽可能地给出更多检验其是否合理的约束条件。

一致性

在一些情况下，不同字段间的数据有相互印证的可能。譬如年龄或出生年月数据与身份证号，就可以相互对照检查数据正误。

总而言之，数据工作考验一个人认真负责的态度，应当尽力去找寻解决方案。

7.2.3 归一化

说到归一化，大家可能会比较陌生，这是由汉语博大精深的历史与文化渊源造成的，因为我们常常想到的与归一相关的成语是“九九归一”，其意思是“绕了不少圈子，最后又还了原”。而我们这里说的归一来自于英文 *normalization*，实质上应该翻译成“规范化”，其本质应该是将多个有量纲的数变成一个没有量纲的数。换言之，就是把绝对的数量改变成相对的数量。

怎么理解上面的绝对变相对，有量纲变无量纲呢？我们尝试这么解释，一个班里有三名同学，身高分别是 100cm、90cm 与 95cm，这里的量纲指的就是单位，即厘米（cm）。当需要变为无量纲的情况，就可以将上述三位同学的身高数据转化成 1.0、0.9 与 0.95。这个时候这三个数没有任何的单位，并不能看出它们到底代表什么，只能知道它们之间的大小关系和比例关系。再换一个例子，这三名同学又进行了一次考试，分别得分 92、98 与 67，其量纲就是分数，而进行归一化（以 100 分为基数）得到 0.92、0.98 与 0.67，无法看出含义。

从上面这个例子可以看出，归一化表象上是对有量纲的数进行处理，而本质上是将原本带有单位的绝对数量变成了没有单位的相对数量，经过这样的变换，我们并不清楚这些数值是什么意思，只知道它们之间的相对大小和比例关系。

那么问题来了，究竟为什么需要归一化，又该如何进行归一化呢？

为何需要归一化

除了去量纲，进行归一化还有一个好处，那就是避免极值问题。例如对于一次统计来说，一个指标统计的是 3，而另外一个指标统计的是 30 000，如果要在同一个图标上进行展示，则几乎看不到 3 这个数据，因为已经被 10 000 倍的比例所稀释了。而如果进行归一化，就可以缩小这样的差距或比例，至于具体采用哪种方式进行比例缩放，而哪种方式又可以实现差距缩小，则要依据不同的情况来定，也就是下文中如何进行归一化的内容了。除了上面两个归一化的原因，还有一个原因，一些算法模型需要将数据进行归一化作为输入，而这部分内容对于数据产品经理来说可能过于深奥，因此不再涉及。

如何进行归一化

这里介绍三种方法。分别是最值归一化、均值方差归一化以及非线性归一化。

最值归一化是使用一组数据中的最大值和最小值进行归一化的策略。这样的方式适用于有限定范围的数据，所采用的公式如下。

$$X_{\text{归一化后}} = (X_{\text{归一化前}} - X_{\text{最小值}}) / (X_{\text{最大值}} - X_{\text{最小值}})$$

例如一组数据中有“10, 9, 8, 7, 6, 5”，将 10 进行归一化后则得到如下结果。

$$X_{10 \text{ 归一化后}} = (10 - 5) / (10 - 5) = 1$$

整组数据归一化之后变成了“1, 0.8, 0.6, 0.4, 0.2, 0”。

均值方差归一化一般适用于没有明显边界的数据，而且最好是正态分布的数据。进行处理的公式如下。

$$X_{\text{归一化后}} = (X_{\text{归一化前}} - \text{均值}) / \text{标准差}$$

这样就可以把服从某种分布的数据变换为标准正态分布数据。

非线性归一化适用于处理极值情况，如展示 10 与 100 000 的情况，当我们采用 log 运算进行非线性归一化的时候，就变成了 $\log(10)=1 < \log(100\ 000)=5$ ，这样两个数在表格上就可以等量齐观地进行展示了。

7.3 聚细沙成佛塔（数据集成）

在大数据时代，我们时常听到一些概念，如打通数据孤岛、链接数据块、横向贯穿数据烟囱等。这些概念虽略有差异，但其都包含同样的含义，即数据分散不如数据汇聚在一起时价值高。

要想把不同孤岛数据汇聚在一起，就必须解决数据的聚变现象。所谓数据聚变，就是当两个数据相遇的时候，会因为重复、冗余以及包含相同意义而被合并成一个数据。就好像两个氢原子在高能的作用下会聚变成为氦原子一样。发生数据聚变的情况有两种，一是指向同一实体，二是出现冗余。

7.3.1 实体识别

说到实体识别，我们首先要说什么是实体。根据百度百科的定义：实体是存在于现实世界中并且可以与其他物体区分开来的物体。

根据上面所说，实体就是名词，也就是说人名、地名、物名都是实体。在计算机领域进行实体识别是一项复杂工作，好在我们并不想弄明白里面的机理，因而只需要清楚在数据清洗的过程中我们需要怎么对待实体即可。

大致说来，我们需要在数据清洗的时候把两个原本不是同一内容的实体区别开，也需要把原本是同一内容的实体整合起来。这就是我们的几项工作。

同名异义

意思是有着相同的名称却包含不同的意义。例如苹果既可以代表手机也可以代表水果。再比如王伟是一个很普通的名字，但是它表示不同的实体。

异名同义

意思是名称虽然不同，但却表示同一个意义。例如我们团队中有一位“涛哥”，名字叫“张涛”，很多场合下我们得知道这是同一个人。又比如“李白”和“李太白”指的是同一个人。又如我们会习惯性地给某个人加上职位性的称谓，如“陈主任”“王博士”“周院长”等。我们要将这些称谓与之真实姓名对应起来。

单位统一

用于描述同一个实体的属性有的时候可能会出现单位不统一的情况，也需要统一起来，如 1200cm 与 1.2m，要知道计算机在进行处理的时候是没有量纲的，要么统一量纲，要么去量纲（归一化）。

ID-Mapping

ID-Mapping 实际上是一个互联网领域的术语，意思是将不同数据库或账号系统中的人对应起来。例如你办了中国移动的手机卡，运营商就会知道你用的是某个手机号，而当你使用今日头条就会留下各种浏览新闻的痕迹，如果现在中国移动要和今日头条合作，那么就得打通两边的数据，第一步就是知道中国移动的张三就是今日头条的张三，这个过程在当下可以通过比照设备的 IMIS 号码来实现，其他的 ID-Mapping 则需要采取不同的策略。归根到底，ID-Mapping 需要的是采用唯一识别号（学号、学校-年级-班级-姓名、设备号等）进行账号的用户匹配。这在大数据强调的数据孤岛问题的解决上有着重要的意义。

7.3.2 冗余性识别

在数据极度丰富的时代，数据量过大已经是被公认的事实。

对于企业来说，数据量大是好事，但似乎又令人苦恼。设想一下，如果给计算机写一个循环程序，向磁盘中不停写入“Hello World”，计算机磁盘不一会就会被塞满，我们能否以此骄傲地说我们拥有了大数据呢？显然不能！这就是数据冗余带来的困扰。另一方面，企业存储的费用是惊人了，往往一个存储服务器起步便是以 10 万计算的，这样的购置价格使得我们更要谨慎、严谨、负责任地对待企业的数据价值，区别有价值与没有价值的数据。

下面从数据冗余性产生的原因以及如何解决数据冗余性两个角度来介绍。

冗余性产生原因

产生数据冗余性大致有两个原因，一个是无意中存多了，一个是故意存多了。

对于无意存多了，举个最简单的例子，QQ 和微信大家应该都用，在 2016 年下半年的统计数据中，双方都声称自己的月活用户有 8 亿多，两个相加就有 16 亿，

这里面必然有重合的，更何况还有一个人有多个账号的情况。多说一句，现在很多厂商生成的覆盖了多少用户的数据，实际上应该是设备数或者账户数，至于这些设备和账户究竟是多少人，这个不得而知。

还有一种情况就是故意存多了。为什么会存在这样的情况呢？主要是为了灾备。灾备，就是为了防止灾害而做的备份。本质上说就是将数据放在一个机器上不安全，为防止被人偷，被黑客黑，被自然灾害弄垮等，于是复制多份。目前通用的做法是将数据备份三份。这些数据放在一起的时候，也要注意去重备份数据。

解决数据冗余性

为了解决数据冗余性，可以从两个方面下手，一个是解决字段冗余，另外一个解决是解决样本冗余。

字段冗余就是指另外一个字段可以从某个字段中推导出来，比如年龄这个字段就可以由身份证或者出生年月这个字段推导出来，利润率这个字段也可以由销售额和利润率这两个字段进行组合运算得到。

而样本冗余就是上述的数据冗余，这个时候需要做的就是识别相同的实体并加以合并。具体的技巧见 ID-Mapping，此处不再赘述。

7.4 换个姿势再来一次（数据变换）

从砂砾中淘出来的金子往往都是碎金，熔合成金块后也不能如我们的意。需要经过锤炼，才能够将其变为配件、饰品等出入商场柜台的商品。对于数据变换来说，主要有两点，一是离散化，二是属性构造。

7.4.1 离散化

离散化是什么意思？我们首先要理解“化”这个字的内在含义。举个例子，“拟人化”是什么意思？其意思就是把非人的内容变成人的样子，所以我们可以总结出这样的一个模式，即“XX 化”就是“把某物变成 XX 的样子”。套用刚才的举例，离散化就是把数据变成离散的样子。

那么究竟怎么理解离散呢？可以用离散的反义词连续来进行对比。不妨让我们看一个例子，请大家思考，年龄到底算不算离散数？年龄的取值可以从 0 一直到 100 多，而且几乎是整数，从某种意义上来看中间并非是连续的，因为很多小数部分都是没有办法取到的，因而可以说它是离散的。但是如果 we 根据人的年龄把人分为幼儿、青年、中年、老年，那么年龄的具体数值看起来就更加连续，至少我们可以说它没有那四个阶段的划分方式离散。因而这个时候可以把年龄看成连续的。由此可知，连续和离散完全取决于我们站在什么样的评判角度上。

离散化总共有四种形式，下面分别介绍。

简单离散

刚才举的将年龄段分成幼儿、青年、中年与老年的例子就是简单离散。即通过定义一些划分规则，将原来连续的数据划分成不同的类别，从而将数据离散化。

分桶后平滑

所谓分桶就是先根据数据的情况设置一些阈值，如收入有 1k, 2k, 3k, ..., 10k, 我们就可以划分为桶 1 (1k~3k), 桶 2 (3k~5k), 桶 3 (5k~7k), 桶 4 (7k~9k), 桶 5 (9k+)。这样原来的十个收入数据就会落到 5 个桶中，在每个桶中可以分别使用落入该桶中所有收入的平均数、中位数以及边界值来替代桶中所有的数值，这三种方式分别称为平均平滑、中值平滑以及边界值平滑。一般来说，我们使用平均平滑较多。

聚类离散

所谓聚类离散就是把相同、相似以及相近的数据进行聚合。采用的是聚类的算法，对每一个聚类的簇进行命名后就是离散的结果。

回归平滑

所谓回归平滑是指，对两个有相关性的变量进行拟合，用拟合线上的数值代替原来的数值，本质上是进行数据噪声的处理。当然很多人也会问，这和数据离散有什么关系呢？确实，直观上这和数据离散没有关系，但是进行回归平滑后的数据已经具备了线性条件，因而可以使用非常明确的边界值对其进行划分，分成两个或者多个。所以让我们姑且也将其放入离散化的行列吧。

7.4.2 属性构造

属性构造是非常抽象的词汇,如何才能让产品经理们理解这个抽象的词语——属性呢?尽管有的产品经理可能会说:“属性,我懂的。”但我还是不放心的,我要和你们说说它的近亲,好让你明白当说到属性的时候究竟是在说什么。

相信除了属性,你还听说过字段,听说过特征,它们其实都是同一个意思。如果你用过 Excel 表格,那么这些属性也好,字段也罢,就是 Excel 表头上的内容。假设你的 Excel 表格存储的是一个学校的学生信息,那么字段(属性,特征)就是姓名、年龄、身高、家庭住址等。

构造属性简单说来有两种方法。

特征工程

所谓特征工程,是把特征提到了工程的高度。工程实施过程中需要人的参与,而且严重依赖人的参与,通过属性与属性的连接,构造新的属性,这就是特征工程。譬如我们已经有属性“长”与“宽”,我们就可以构造“面积”这个属性。我们有每天的上网时间这个属性,就可以构造一周上网时间这个属性,也可以构造工作日上网时间属性以及周末上网时间属性。

随意构造后筛选

随意构造与人工构造的区别是,人工构造的属性往往是有意义的,如上面所列举的,但随意构造的属性往往没有意义,也就是说任意两个或三个乃至更多的属性都可以组合在一起,加减乘除都没有问题。那么构造了这么多的属性,怎么判断该使用哪个属性不该使用哪个属性呢?这里就可以用到 XGBoost 工具来进行筛选。由于这里已经涉及较为深奥的机器学习知识,超出了本书介绍的范畴,因此不再展开。

7.5 少即是美(数据规约)

“山不在高,有仙则名;水不在深,有龙则灵”,数据不在多,有价值就行。大数据时代,我们忧愁的不是数据太少,而是太多。数据规约提供了一种针对数据过多的解决方案。按照这个步骤进行,数据也会变得不再“大”了。

7.5.1 特征规约

规约这个词听起来比较奇怪，理解它的意思可能有些难度。规约，规是规定的规，约是约束的约，可见规约有限制的意思。从说文解字的角度来讲，限制有减少范围的意思。因而我们可以认为特征规约就是在减少特征。

规约的基本原则就是不产生信息丢失，不影响分析结果。然而这两个评判准则是根据实际的分析任务来定的。进行特征规约有两种形式，一种称为立方体规约，另外一种称为维度规约。下面分别介绍。

立方体规约

要弄清楚什么是立方体规约，首先要明白什么是立方体。我们听说过正方体，正方体是一个三维的立方体，这就意味着特征只有三种，如区域、学生数、产品类型。而对于多维的立方体，则有多个特征。进行立方体规约，就是将 N 维的立方体变成 $N-1$ 维的立方体。

那么少的一个维度为什么需要去掉呢？实际上是为了累计计算，就三维立方体（中间的数据是学生数）来说，我们把区域、学生年级、产品类型变化为区域、学生年级，实际上就是统计了每个区域各个年级内使用所有产品的学生数。根据累计统计目的不同，我们应该减少不同的维度，例如为了统计每个区域内使用各个产品的学生数目就应该减少学生年级这个维度，进而变为区域、产品类型。

维度规约

维度规约就是去除与业务分析无关的属性。例如我们要分析和学生学业有关的影响因素，维度中如果有老师的身高这类不相关的因素，便可以去除。这是根据业务逻辑进行的去除。我们还可以根据这个属性的重要程度来评估是否需要去除，而属性的重要程度可以采用决策树来评估（算法具体后面介绍），凡是不在决策树上的属性都可以尝试去除。

所以本质上，立方体和维度规约都是在减少特征的数目，而其区别则在于我们的目的不同。立方体规约是为了累计计算，而维度规约是为了去除无关特征。

7.5.2 样本规约

一张表格应该有自己的修养。表格的自我修养应该包括“行”的修养与“列”的修养。如果说特征规约是列的修养，那么样本的规约就是行的修养。因为特征决定了列数，而样本数量则决定了行数。

同样地，进行样本规约的过程应该做到不产生信息丢失，且不影响分析结果。总结来说，总共有三种类型的样本规约，分别是去除冗余、抽样与压缩编码。

去除冗余

在样本中必然会有相同数据存放两遍的情况，这时候，如果不是为了数据分析的必要而做的冗余存储，那么就需要去掉冗余的部分，保证相同的数据仅保留一份在数据库中，这就称为去除冗余，这样能够减少样本的数目。

抽样

所谓抽样，就是不选取所有的样本数据，而是从中抽出一些样本。抽样又有N多种方式。

首先是有放回抽样，这种抽样方式使得相同的数据可能会被再次抽到。其过程是从一整个表中抽出一行，在一个本子上记录下来，然后再从这张表中抽出一行，再记录在刚才的本子上，周而复始，依次进行。

其次是不放回抽样，这样的过程使得获取的数据一定是原始数据的一个子集。其过程与上述有放回抽样的不同之处在于，当从原始表中抽出了一行之后，不仅要在一个本子上记录下来，而且还要在原始表中把这行数据抹掉，使得下次随机抽样的时候不会再抽到这行数据。

接着是分层抽样。抽样的目的是为了选取样本中的代表，有点类似于选举会议代表的过程。上述有放回抽样与无放回抽样都假设所有候选人的中举机会是均等的，但是事实并非如此。譬如说我们需要考虑男女比例的分配，少数民族群众的比例，无党派人士的情况甚至基层贫困群众的情况等，这些因素使得不同人群被选举机会的均等程度不同。因而将人群划分成若干群体，并在不同的群体中使用随机抽样的过程就称为分层抽样。

最后是聚类抽样，如果说刚才的分层抽样是人为划分群体的话，那么聚类抽

样就是用计算机来判断哪些人该聚成一群，其余的和分层抽样一样。

压缩编码

上述减少数据量的过程本质上是减少了数据的样本数，而另外一种方式是将数据进行压缩再存储，例如采用 01 的稀疏编码表示。这个技巧太过技术化，产品经理不需要掌握，只需要知道有这个技巧即可。

第 8 章

面向产品经理的统计分析

- 8.1 说有信息量的话（非时序数据的统计量） 144
 - 8.1.1 集中趋势 145
 - 8.1.2 离散趋势 146
 - 8.1.3 数据分布 148
- 8.2 股票指数是什么（时序数据的统计量） 148
 - 8.2.1 “三比” 149
 - 8.2.2 股票指数 150
- 8.3 男女真的有别吗（分类数据的统计量） 152
 - 8.3.1 卡方是什么 152
 - 8.3.2 卡方怎么算 153
- 8.4 相关性不是因果性（连续数据的统计量） 156
 - 8.4.1 Pearson..... 156
 - 8.4.2 Spearman 157
 - 8.4.3 Kendall..... 158
- 8.5 数据不能承受之“熵” 159
 - 8.5.1 物理中的“熵” 159
 - 8.5.2 信息中的“熵” 160

经过清洗的数据已经整齐地摆放在面前，它们整齐划一，仿佛是等待检阅的卫兵。是时候对它们进行一些基本的队列训练了。在数据分析领域，我们称这些基本的训练为统计分析。

学术领域对数据有严格的区分，最简单的两个角度分别是“定性”与“定量”。例如我们说一个人是好或坏，男或女就是一种定性的说法。而如果对其身高体重、血压心率进行测量，则得到了定量的结果。企业中的 KPI 便是一种定量的考核方式，与之相对的便是对一个人的定性评价，如勤奋、偷懒等。这里并不打算严格使用学术上的数据定义方式来介绍统计分析的基本方法，我们可以把每一个数据项看成队伍中的人，然后依据此来介绍四种最常见的数据类型，以及在这些类型数据上能够进行的操作。

队伍有种类之分，数据也有类别之分。一些数据无序地堆放在一起，仿佛是随机聚合，又随机散开，没有谁在谁的前面，好比是散兵游勇，这样的数据是“无序”数据。经过训练的队伍有整齐划一的步伐，方阵中谁在谁的前面是一定的，很多情况下是按照先来后到的时间顺序排列，这样的数据是“有序”的，仿佛是正规军的步兵师。当引入机械化和信息化手段后，部队进一步分化成不同的类别，如炮兵、步兵、装甲兵部队等，数据中的每个样本也依据情况被分成不同的离散类别，如男、女，这样的数据是分类数据。与分类数据相对应的是连续数据，如睡眠时长、饭量这些指标，可以使用分钟、小时与重量、卡路里来衡量，这就好比具备多种技能而无法相互区隔的全能型特种部队。

8.1 说有信息量的话（非时序数据的统计量）

散兵游勇型的数据学名叫“无序”数据，准确说应该叫作非时序数据。我们大多数情况下遇到的数据都是这样的数据。例如一个班级的学生成绩，某单位的人员收入，医院某科室的病患检测指标等。

很多人会不解，学生成绩或者收入水平这样的指标有高有低，明明可以排大小。再譬如体检时测量身高这个指标，总可以按照身高从高到低有秩序地排成一列，为什么说是没有次序的呢？诚然，这里使用的“无序”需要添加一个引号。因为这些数据并非没有次序，而是没有时序。

对于这样的数据我们究竟该如何洞察它们呢？这里介绍三种统计量，分别用于描述数据的集中趋势、离散趋势以及整体分布。

掌握统计量的目的是为了变定性为定量，这样就不会在别人询问数据质量如何的时候，使用诸如“数据质量不行”这样的话去搪塞，而是能够更加专业地给出一些有信息量的回答。

8.1.1 集中趋势

数据的集中趋势就好比数据所具有的集体性指标。通过了解数据的集中趋势，我们可以对整体数据情况有个把握。我们熟悉的平均值、众数、中位数以及四分位数都是这种性质的统计量。

平均数

有一个有趣的故事是用来描述平均数的。故事是这样的，一个房间中坐了三个人，分别拥有 80 万元、100 万元与 120 万元的资产，这时这个房间中每个人的平均资产是 100 万元。通过平均值这样的统计量，我们可以大致了解一个群体中人们的整体状况，这样看来还算准确。但这个时候一个富豪推门而入，这个富豪拥有 10 亿的资产，瞬间拉高了这个房间中人们的平均资产，当我们再使用平均值进行估算的时候，发现人均资产瞬间超过了 2.5 亿。如果不在事先知道之前三个人的整体状况，单看这一个指标，会误以为房间中的每个人都是富豪，而实际上却并非如此。因而平均数只有在大家都差不多的情况下进行定量判别才是较为准确，且有意义的。

众数

众数顾名思义就是数量最多的那个数。我们在电商网站上购物往往会选择按热度或者销售数量排列，这两种情况就是分别针对商品的访问量和销售数量进行比较，排名第一位的就是有最多人选择的商品，故而在进行排序选择的时候就会排在第一位了。使用众数进行判别会不会出现什么问题呢？答案是一定的。在使用众数进行判断的时候容易陷入媚俗、民粹与从众的陷阱，大家一定要注意。

中位数

既然平均数与众数不能反映数据的真实情况，总会受到极端情况的影响，那么有没有一个数字可以部分消除极端个别数据的影响呢？中位数就是这样的一个统计量。将一些数字从大到小排列，无论是定性的（上、中、下）还是定量的（如身高、体重等），中位数就是位于队列正中间的那个数字。以一群人的收入为例，无论参与排列的有无个别巨富之人，中位数的波动往往不大。

四分位数

四分位数是中位数的“亲戚”，和中位数一样，求得四分位数需要先将数据按照一定规则排序（如按照从小到大排序）。四分位数实际上是三个数，中位数是其中一个，又称为第二分位数。第一分位数位于中位数左边的正中间，第三分位数则位于中位数右边的正中间。通过分位数，可以把排序好的数据划分成等距的一段一段，每一段中数据量相同，分位数就是每段首尾相接的那个数。分位数求得越多，这个样本的整体情况就了解得越清楚。

8.1.2 离散趋势

如果说集中趋势是整体情况，离散趋势则反映了更多的个体特例情况。掌握离散趋势的统计量，就等于抓住了数据的细节。这里可以使用的统计量包括方差、标准差、变异系数、极差以及最值。

方差

方差的英文是 `variance`，这个单词由两个词根构成，分别是 `vari-`与`-ance`。其中`-ance` 表示某种性质与状态，如 `difference`（元音字母可以相互替代）与 `importance`。而 `vari-`则是一个拉丁语词根，表示变化（`change`）。这个词对于程序员来说再熟悉不过了，用于定义变量的关键字就是 `var`。由此可见，方差（`variance`）一词是用来表示数据中成员间不同及相互变化程度的。既然是计算不同，求差即可，为何又要在前面加上方这个字呢？这很好理解，求差就会有正负数，如果仅仅使用差的累计来计算数据间变化和差异的程度，就会造成相互抵消。一个简便的办法便是对于求得的差值进行平方后再累计，这也就是方差一词的来历。

标准差

标准差是方差的“近亲”，是对方差求平方根后的得数。可是为什么有了方差还需要标准差呢？原因就在于量纲。量纲一词我们之前已经介绍过了，这里以身高为例，假设群体的身高平均值是 170cm，某人的身高是 172cm，我们可以说该人身高比平均值高 2cm。这 2cm 是差值，但在计算方差的时候需要对这可怜的 2cm 进行平方运算，得到的量纲也随之变成了 cm^2 。虽然我们可以用方差来表示个体间变化和不同的程度，但是很难解释 cm^2 在这里表示什么，难不成是面积？显然不对！但是当对方差求平方根后，量纲也随之变为 cm，于是标准差在现实世界中又有了可以解释的意义。虽然标准差在计算的时候由方差求得，但其实标准差的诞生要比方差早了将近 20 年。1894 年英国统计学家卡尔·皮尔逊（Karl Pearson）提出了这个概念，24 年后的 1918 年，同为英国统计学家的罗纳德·费雪（Ronald Aylmer Fisher）提出了方差的概念。

变异系数

平均值看整体，标准差看差别，相信大家已经很熟悉了。我们常用标准差来表示一组数据的稳定性。例如在工业生产中制造一种材料，通过测量其纯度来观察生产制造的稳定性。直观上，方差越小，表示制造系统越加稳定。但是有的时候我们需要比较生产不同材料的生产线制造工艺，这就比较麻烦了。例如，对于生产 A 来说，其纯度要求可能是 99%（平均值），对于生产 B 来说，其纯度要求是 95%。在两个生产线上测量得到的方差又不同，这该如何是好？一种最为简单的方法就是用比例这样的相对值替代（标准）差这个具有量纲的绝对数值，即了解偏差的大小占其平均值的比例是多是少，这便是变异系数的定义。

最值

最值对于我们来说再熟悉不过了，分为最大值（极大值）和最小值（极小值）两种。最值表现了一个排好序的数据队列其两端的数据情况，是用一个个体来表示群体某种细节特征的统计量。

极差

单从名称上就知道，极差是极大值和极小值的差值。它还有一个名称叫作全距，英文是 range，即全部数据的距离和范围。range 有山脉的实意，例如位于澳大利亚珀斯（Perth）附近的大岭山脉（Darling Range）。由于山脉绵延起伏，因而

有跨越一定区域的范围之感，故而引申出了范围的意义，用在数据分析上便是全距的概念。

8.1.3 数据分布

数据分布相对来说较为高级，一般来说，探讨一些深层次问题时可能会用到。这里的分布也是一个比较宽泛的概念，囊括了除分布概念外的信息熵、偏度、峰度等。当然本质上还是指各种分布的名称，譬如大家耳熟能详的高斯分布（正态分布）、指数分布、二项分布、beta分布、Dirichlet分布等。相信非理科背景的产品经理看到这应该有些懵了，告诉你个好消息，你只需要知道数据分布需要被考量，且只需要知道高斯分布与指数分布即可。

高斯分布

简单来说，高斯分布就是真实世界的分布，中间大两头小是它的特性，由于其曲线像一口倒扣的钟，所以也称为钟形曲线。以班级的考试成绩为例，特别优秀的学生是少数，特别差的学生也极少，大部分学生处于中游，因而这个班级的成绩分布就是高斯分布。高斯分布还有一个别名叫作正态/常态分布，因为自然界中正常状态下这样的分布很常见。我们对于一个班级下结论说其考试成绩服从正态分布是武断的，在学术界看来，最多只能叫偏态分布。因为由班级学生成绩围成的钟形曲线并不对称，因而只能是偏向一侧的偏态分布。

指数分布

既然叫指数分布，自然就是指其形状与指数函数的图像类似。但对于大多数文科背景的产品经理来说，指数函数是什么可能都已经忘记了。指数函数曲线从坐标系第一象限的左上角向下弯曲一直延伸到第一象限的右下角，满足这种分布的曲线就是近似于指数分布的曲线。

8.2 股票指数是什么（时序数据的统计量）

无序是一种常见的状态，如物理学上的布朗运动，原子核外层电子的运动与跃迁，街道上熙熙攘攘的人群，以及城市中漫天飞舞的灰尘。而从无序到有序则

是人们一直追求的，因为有序能够带给人们确定性与安定感。

那么对于数据来说，有序是什么样的一种状态呢？我们可以举几个例子来看看什么是有序的数据。

第一个例子来自于大型购物超市。有过消费经验的人一定记得，在每次结账完成后都会获得一张小票，小票上会记录订单的流水号、购买的商品种类和数量、单品价格以及总价、购物时间等信息。如果是在电商平台购物，则可以获得电子版的消费账单。在超市或者电商网站的后台，这些数据其实被悉数记下，用于分析用户的消费习惯。对于一个用户来说，如果购物的次数不止一次，那么每次的消费数据叠加在一起就形成了一个先后的关系，我们称这样的数据为时序事务型数据。

第二个例子来自于医学界。在 1990 年的时候，人类基因组计划（Human Genome Project）正式启动，基因测序的工作在那时异常复杂与耗时。在医学愈发昌明的今天，进行基因检测已经非常简单，我们可以测得组成基因的碱基对的排列顺序。打个不太恰当的比方，碱基对就像是火腿肠，测序后得到的数据就是首尾相接的火腿肠。碱基对总共有 4 种类型，正是这些碱基对的排列顺序不同造成了人与人之间的千差万别。这样的数据我们称之为序列数据，序列数据中的数据位置不可变换，否则就构成了一个新的序列。

第三个例子来自于气象领域。以前几年最常发生的雾霾现象为例，几乎在所有城市，都会常态化地检测 PM2.5 水平。对于一个城市来说，以时间为轴，就可以得到该城市雾霾情况的波动水平。这样的数据就是时间序列数据，由于时间是不可能相互颠倒的，因而时间序列数据也最为有序。如果把时间序列数据叠加在城市的经纬度信息上，这样的数据就变成了时空数据。

那么对于时序型数据，我们可以使用哪些统计量来进行探究呢？不妨以股票价格这一典型时序型数据为例，介绍三个指标，分别是同比、环比、定比。

8.2.1 “三比”

同比

所谓同比是指和往年的同期进行比较，譬如要探究某微信公众号 2017 年 9 月

的数据状况，使用同比时，就可以与 2016 年 9 月的数据进行比较。可以比较阅读量，可以比较评论数，也可以比较转发和收藏量等。再譬如股票的价格在一季度最高值是多少，成交多少，要计算同比的时候，我们就要和去年一季度的这两个指标进行比较。同比上升还是下降就取决于两个数据的高低。

环比

环比与同比的不同就在一个“环”字，这里的环可以有两重理解。一方面某个指标在一年 4 个季度 12 个月的数据构成一个圆环，我们将圆环上相邻的两个数据进行比较，称为环比；另一方面，环还有周期的意思，扩展一下概念就可以定义为将相邻两个周期内的数据进行比较，这个周期可以是一周、一个月、一个季度等。仍然以股票为例，同比比的是当年 9 月与去年 9 月，而环比则是选取当年 9 月与当年 8 月进行对比。刚才我们选择的时间跨度是一个月，我们还可以将时间跨度变大为一个季度，或者变小为一周或一天。

定比

定比就更容易理解了，所谓“定”就是和某一个固定时期的数值进行比较。很多人会疑惑，总是看过去有什么意义呢？环比和同比就是看近期的，尽管这个近期最多也就是一两年，但总是要相较定比所参照的某个 50 多年前的数据要可靠吧？大家考虑的并非没有意义，然而当我们深入考虑比值的含义时就会发现问题。比值由分子和分母组成。对于同比和环比来说，不同的时间，分母和分子同时在变，这使得我们很难有一个标准来衡量两个不同时期数据的变化。使用定比就是为了确定分母，即进行比较的过程中有一个量是固定的，因而可以使用这种方式来看到数据的波动趋势。事实上，股票市场正是利用定比最多的地方，并产生了股票指数这样的概念。

8.2.2 股票指数

这里介绍的股票指数，算是定比的拓展知识。股票市场中有两个概念，一个是基期，一个是报告期。所谓基期指的是最开始的时间点，也就是基准时间点，而报告期则指的是当下的时间点。

那么股票指数究竟是怎么计算的呢？在过去某个时刻，我们从市场中选几支

股票，计算它们的价值。然后在当前时间选取几支有代表性的股票再计算价值，用当前的价值比过去的价值，再乘以一个基数就得到了股票指数。

如果觉得刚才的一段话比较抽象，我们可以举个简单的例子。例如在很多年前刚开始有股票市场的时候，我们选取了 A、B、C 三家公司的股票价值来衡量整个股票市场的价值，不妨记录为 ABC。等到现在这个时候，对于同一个股票市场，我们选取 A、C、F 三家公司的股票价值来衡量整个股票市场的价值，将 ACF 与 ABC 作比，然后乘以一个基准的数（譬如 100），就得到了股票指数。我们通常听到的纳斯达克指数、沪深指数就是这么计算的。可是为什么两次衡量所采用的公司不同呢？一方面是由于有的公司退市或者倒闭而没有继续存在市场中，另一方面是由于产业与行业在发展，不能仅仅依靠以前的单一行业公司来评估整个股市的价值，因而需要引入多元化的公司类型。

很多人可能会继续发问，难道每支股票都是平等的吗？举个极端的例子，譬如一个公司发行了 1 万股，而另一个公司发行了 1000 万股，它们的价值能一样吗？也许还会有人质疑，难道这样的股票数量不是已经计算在市场价格中了吗？毕竟这个股票的市场总价值等于发行量与单价的乘积呀，那么究竟为什么需要再次区分不同的股票呢？我们或许可以这样解释，发行数量一般来说不会轻易变动，变动的是价格，而价格的变动尽管有剧烈波动的可能，但是毕竟还是小幅度的，更何况多支股票交织在一起有对冲的可能，因而整体的市场价值不会有太大波动。如果不发生变动，又违背了指数发明的最初意义，因为这个数就是用来进行指引，让大家知道趋势的。所以这个数应该变动更加明显，故而还应该在股票总价值前加上一个权重。

在加权重这个问题上，有两派学者，一派认为应使用过去时刻的发行量或者成交量作为权值，而另一派认为应使用当前的发行量和成交量作为权值。前者称为拉斯拜尔指数，后者称为派许指数。目前世界上大多数股票指数都是派许指数。

在基准点设置这个问题上，各个股票交易机构也有不同的做法，有的设置为 100，有的设置为 50，还有的设置为 10。这也就不难理解为什么有的地方股票指数为好几万，而国内总在几千徘徊了。下面列举一些股票指数给大家作为补充学习的材料。

表 8-1 全球股票指数及其信息

股票指数名称	开市时间	所属国家	基准点
道琼斯指数	1928.10.1	美国	100
富时(FT)指数	1935.7.1	英国	100
标准普尔指数	1943	美国	10
纽交所指数	1965.12.31	美国	50
日经指数	1950.9	日本	176.2
恒生指数	1964.7.31	香港	100
上证股指	1990.12.19	上海	100
深证股指	1991.4.3	深圳	1000

其他股票指数还包括了上证 180，沪深 500，央视财经 50 等。

8.3 男女真的有别吗（分类数据的统计量）

我想搞教育的人应该都听说过培生教育集团吧？即使你不是搞教育的，但是你学过 GRE 或者 TOEFL，那么应该买过朗文公司的英语教辅吧！

你也许困惑，这一章讲的是对分类数据的统计分析，与培生教育集团什么关系呢？卡尔·皮尔逊（又一翻译为培生），全名是 Karl Pearson，我们在之前介绍标准差的时候提到过他。这位统计学家创建了培生教育集团。

其实早在培生之前，就有两位科学家致力于统计学研究，一位是高尔顿（Galton），另一位是韦尔登（Weldon）。

因为受到高和韦两位大师的影响，Pearson 致力于推动统计在生物学上的应用。Pearson 很努力，很多统计学上的概念都是由他发明的。譬如标准差、成分分析以及卡方检验等。

8.3.1 卡方是什么

卡方是一个很别扭的词汇，实际上是希腊字母 X，可以认为卡方是 X 的平方。另外希腊字母对于卡方的标音是 chi，其中 ch 发 k 的音，所以叫卡方就不足为怪了。

我们可以先提出一个问题，也是平常生活中经常遇到的问题。究竟如何判断两者是否相关？例如，我们会问，是否家长是大学生，孩子就一定是大学生呢？这实际上问的是家长的学历与孩子学历的相关性。我们还会问，是否作业多孩子的学习成绩就一定好呢？这实际上是问作业量与学习成绩的相关性。我们还会问是否只有女生喜欢化妆呢？这实际上是问性别和化妆是否有相关性。

而上述所有的相关性问题都可以抽象成两个因素之间的关联性问题。但是要注意，因素和因素是不一样的。譬如作业量与学习成绩是可以取很多个值的，但是性别，是否化妆往往是二选一的。我们称前面数不胜数数据是连续型数据，而后面二选一或者几选一的数据之间仿佛拉开了很长的距离，可以数清楚，它们彼此是相互散开的，因而称为离散型数据。又因为离散型数据是有固定类别的，因而我们可以说它们是分类型数据。

卡方就是用来检查分类型数据间相关性的统计量。

8.3.2 卡方怎么算

让我们以网络上最为流行的一个例子来介绍卡方怎么计算。这个例子讨论了男女性别与是否化妆的相关影响。

且让我们抛开数据分析，先用经验来判断一下，化妆这件事对于男女生而言，选择是否有差别呢？是否女生更爱化妆？如果是，我们就可以根据化妆与否来判断男女的性别，至少很大概率上可以判断出来。如果不是，那么就只能说性别和化妆与否并没有很强的关联性。相信大多数人会认为化妆这件事与性别是相关的，究竟情况在大众群体中是怎么样的呢？

下面利用卡方来解决这一个问题。我们走上街头进行数据统计，随机走访 200 人，其中男生 100 人，女生 100 人，并统计他们化妆与否的情况，得到了表 8-2。

表 8-2 男女化妆情况统计

	男	女	合计
化妆	15	95	110
不化妆	85	5	90
合计	100	100	200

拿到这样一张数据统计表格后应该怎么看呢？跃跃欲试的人直奔主题而去，

他们观察到女性化妆的比例远高于男性化妆的比例，于是得出结论：显然女性更加热衷于化妆，于是化妆与否的确与性别有关联。

实际上我们并不能这么简单地判别，我们可以根据化妆与不化妆的人数合计，并依据男女生的比重标出来这样的情况，即如果我并没有观察到男生和女生化妆与否的人数，那么仅仅根据化妆的 110 人与不化妆的 90 人，应该猜测男女生化妆与否的人数是多少呢？如表 8-3 所示。

表 8-3 男女化妆情况估计数据

	男	女	合计
化妆	55	55	110
不化妆	45	45	90
合计	100	100	200

也就是说，在没有任何确凿证据的情况下，我们只能根据男女 1:1 的证据，以及化妆的人有 110 人，推测男女化妆的人数是相等的。不化妆的人数是 90，我们也只能得出男女各 45 人。采用这种方式进行判断，等于假设了一种理想的情况，即男女性别并不会影响化妆与否，性别与化妆这件事是不相关的，性别与化妆这两件事是独立的。如果要想证明它们相关，就要拿事实说话，推翻这个结论。

这个事实就是卡方 (X^2)。第一张表中的调查数据就是实际的情况，当我们把它与理想的状况进行比较的时候，会发现差异的存在。这个差异简单说来，就是原本应该填写 55 的地方变成了 15 和 95，而原本 45 的位置被 85 和 5 代替。实际数字与理想状况的数字愈加吻合，表明不相关的结论越正确。而差别越大，则表明不相关的假设站不住脚。

那怎么计算差别呢？最简单的计算办法就是求差值。如下所示。

$$95-55; 15-55; 85-45; 5-45;$$

如果要对上述差值进行累计就会发生和方差一样的情况，正负相互消除了。因此可以加上平方，使得差值得以累计，这也是卡方名字的来历。修正后的公式如下所示。

$$(95-55)^2+(15-55)^2+(85-45)^2+(5-45)^2$$

看似到这一步已经解决了问题，但实际上由于具体情况不同，上述计算的数值大小会依据不同的场景而有很大的差别。一个最显著的问题是，如果现在调查的样本数量扩大为 200 万人，即便实际调查的数值与理想状况很接近，例如理想状况为男女化妆人数各 55 万人，实际状况为 56 万人，也会因为差值为 1 万，再经放大后获得比上述公式更大的差值。因此绝对的差值累计不能说明问题，我们应该使用比例来进行说明，即实际状况比理想状况偏离的数值占理想状况的百分比。于是公式进一步改动如下。

$$\frac{(95-55)^2}{55} + \frac{(15-55)^2}{55} + \frac{(85-45)^2}{45} + \frac{(5-45)^2}{45} = 129.3$$

总算可以覆盖所有出现的异常状况了，可是最终的卡方值 129.3 是什么意思呢？凭借这个数难道就可以衡量性别和化妆与否的相关性了吗？究竟超过 100 是相关，还是超过 50 是相关呢？

Pearson 给出了一个表格，我们只需要将计算得到的数值与表格中的数值进行比较即可。如下所示。

表 8-4 卡方值与概率对照表

概率	0.50	0.40	0.25	0.15	0.10	0.05	0.025	0.01	0.005	0.001
卡方	0.455	0.708	1.323	2.072	2.706	3.841	5.024	6.635	7.879	10.828

这个表格中的概率是什么意思呢？根据上面提到的，卡方值越小则表明越不相关。那么这个表越往右侧卡方值越大，概率却越小，表明卡方值取某个值的时候，关于“不相关”的假设成立的对应概率。带入刚才的实际情况，129.3 远大于 10.828，即意味着“性别与化妆与否不相关”的论断正确率低于 0.1%，即论断不对，那实际情况就是，性别与化妆与否是相关的。

可是为什么要假设两者不相关/相互独立/没有差异，而不是先假设它们相关/相互不独立/有差异呢？一方面是出于严谨的考虑，在一切都没有被证明前我们倾向于保守；另一方面，和当今法律中的无罪假设一样，证明存在比不存在容易，可以节约资源，提高效率。因为证明存在（是真的）只需要举例即可，而证明不存在（是假的）则很难。

这就是 Pearson 对于判断分类数据相关性的贡献，希望你已经大致了解了。除此之外的 t 检验、U 检验、F 检验与卡方检验一样，只不过判断的场景不同。

8.4 相关性不是因果性（连续数据的统计量）

我们时常能够听到正相关和负相关的说法，那么究竟怎么样判别两组数据是正相关还是负相关呢？是否还可以采用之前说到的卡方呢？显然不行，因为卡方对应的概率最小值是 0，而最大值是 1，顶多只能决定是相关还是不相关，根本覆盖不到正负这么细节的情况；其次，我们之前也说过，卡方是用在离散型数据上的，对于连续型数据是不适用的。

不用焦躁，解决方法马上奉上。这里主要介绍以下三种方法，这三种方法的相似点都是使用一个值表示数据是否相关，而且这个值取值都在-1 到 1 之间，1 表示正相关，-1 表示负相关。但是解决思路各不相同，下面一一介绍。

8.4.1 Pearson

皮尔逊（Pearson）已经被我们多次提及，他的最大贡献是为现代统计学打下了基础。很多大家熟知的统计学名词，如标准差、卡方检验等也都是他提出的。

Pearson 出生于 1857 年的英国伦敦，他是一名数学家，也是一名生物统计学家。1884 年，他进入伦敦大学任教。

就在其任教后不久，统计学界的又一丰碑人物高尔顿出版了《自然遗传》，书中提出了回归的相关概念，并阐述了其用于研究生物特性的价值。Pearson 对这个理论着了迷，于是决定为统计学全力以赴。就这样，Pearson 追随者高尔顿与韦尔登的足迹，为了推广统计在生物学中的应用，与二位前辈一起创办了《生物统计》（Biometrika）杂志，并提出了 Pearson 相关系数。

回到主题上，对于 Pearson 来说，他求解的是相关系数。整个计算过程与计算向量夹角余弦的公式差不多。例如，对于向量 $x = (1, 2)$ ， $y = (2, 3)$ ，计算余弦值可以使用如下的公式。

$$\cos = \frac{1 \times 2 + 2 \times 3}{\sqrt{1^2 + 2^2} + \sqrt{2^2 + 3^2}}$$

对于 Pearson 相关系数的计算来说，只不过要在计算前将每组数据减去其均值（也称为数据居中），如下所示。

$$\text{Pearson} = \frac{(1-1.5) \times (2-2.5) + (2-1.5) \times (3-2.5)}{\sqrt{(1-1.5)^2 + (2-1.5)^2} \times \sqrt{(2-2.5)^2 + (3-2.5)^2}}$$

众所周知，夹角余弦的取值范围恰好是-1 到 1，因而 Pearson 相关系数在公式上与之相似也就不足为奇了。

既然有了夹角余弦公式，还需要 Pearson 相关系数的原因就在于，Pearson 进行了数据居中操作。这个过程对于推荐系统来说十分重要。例如我们在使用点评类网站的时候，会给商家或者商品打分。如果是一个苛刻的消费者，则打分普遍偏低，如果是一个标准松懈的消费者，则打分普遍偏高。为了消除个人偏好对数据高低的影响，就需要在各个数据中减去他们各自打分数据的均值，从而得到他们对于某个商家看法高于其心理均值程度的数据，这也就是 Pearson 进行数据居中的目的。

8.4.2 Spearman

Spearman 全名是查尔斯·爱德华·斯皮尔曼 (Charles Edward Spearman)，主要的成就领域在于心理学和逻辑学，在 1923 年到 1926 年期间担任了英国心理学会的主席。

由于在进行心理学实验的过程中需要了解影响因素与结果的相关程度，于是 Spearman 在心理学统计方面做了大量的工作，并提出了我们即将介绍的 Spearman 相关系数。

对于 Spearman 来说，他求解的是 Spearman Rank，翻译成中文叫作“斯皮尔曼的秩”。秩的意思似乎有些难懂，但是当你用其组词，你会组出“秩序”这个词，因此这个词也暴露了其解决问题之道，即使用数据的排列顺序来获得最终的计算结果。

Spearman 将数据重新按照由大到小的顺序进行排列，然后计算两组数据间排位的绝对差值，最终使用一个固定公式来进行计算，把结果限定在-1 到 1 之间。假设有两组数据，举例说明如下。

表 8-5 Spearman Rank 计算示例数据

数据 1	数据 2	数据 1 排位	数据 2 排位	排位差
170	180	3	2	1
150	165	5	5	0
210	190	1	1	0
180	168	2	4	-2
160	172	4	3	1

我们已经按照 Spearman 的要求对数据 1 和数据 2 中的数据从大到小进行了排列,记录了相应位置上的数字在整体中的排位(如数据 1 排位与数据 2 排位所示)。并基于两组数据的排位计算了差值,如排位差所示。

根据 Spearman 总结的公式就可以计算出 Spearman Rank。这个公式是一个形式复杂但结果简单的庞然大物,在此就不写出来了,其中只需要两个量,一个是两组数据的个数,在这个例子中是 5,另一个是排位差的平方和,在这个例子中是 6。

8.4.3 Kendall

Kendall 的全名是莫瑞斯·乔治·肯德尔 (Maurice George Kendall)。出生于 1907 年,后进入剑桥大学圣约翰学院学习。毕业后,他进入了英国农业部工作,在那里,他开始对统计学产生兴趣,并致力于使用统计学去解决农业中的问题。

Kendall 的工作与 Spearman 很像,其最终提出的用来描述相关性的数值被称为 Kendall Rank,即“肯德尔的秩”,因而也不言而喻地采用了数据的排列顺序来获得最终的计算结果。

但是其和 Spearman Rank 的区别在于, Spearman 强调的是两组数据在所在组中位置次序的吻合程度,而 Kendall 强调的则是,将固定一组数据从大到小排列,查看其对另外一组数据位置的影响。我们仍以上表中的数据为例,不妨将数据 1 从大到小排列,数据 2 也对应调整了位置,从而使得数据 1 与数据 2 还是一一对应。得到下面的表格。

表 8-6 Kendall Rank 秩计算示例数据

数据 1	数据 2	数据 1 排位	数据 2 排位	同序对数目
210	190	1	1	4
180	168	2	4	1
170	180	3	2	2
160	172	4	3	1
150	165	5	4	0

在上述表格中出现了一个概念：同序对。该怎么理解这个概念呢？例如将数据 1 排序后，对于一对数据，如（210，190），如果随便拿另一对数据比较，如（160，172），210>160 且 190>172，就称这是一个同序对。否则对于像（180，168）与（170，180）来说，180>170 但 168<180，这就不能算是一个同序对。在上面的表格中，210 与 190 分别是数据 1 与数据 2 中的最大值，因此，（210，190）共有四组同序对。

同样地，当获得了同序对数目之后，只需要知道数据的个数，就可以得出 Kendall Rank 了。公式在此也略去不谈了。

8.5 数据不能承受之“熵”

在前面介绍数据分布的时候，我们谈到了一个概念“熵”。但是这个熵并不是我们说的智商、情商与财商，而是一种物理学的概念，由于信息学是由硅（Si）堆叠起来的，因而熵也自然而然与信息学相互联系了起来。

8.5.1 物理中的“熵”

如果想要聊清楚熵的概念，还得从大家耳熟能详的“士力架”开始。士力架与熵有什么样的关系呢？士力架包装上清楚标明了卡路里，也就是我们平常说到的食物热量。从士力架的例子可以看出，热量和能量好像从本质上没有太大区别。一般来说，热量高的食物能量也高，而能量高的食物也有相当高的热量。严谨来说，不能认为两者是相等的，但是至少从科普的角度来说，两者可以近似等价。

时间来到 1923 年，量子力学之父普朗克（现在物理学上有一个概念叫作普朗克常量，就是以他的名字命名的）在南京讲学，说到热力学第二定律的时候用到了 Entropy 这个词语（熵的英文称呼），粗浅来说 $\text{Entropy} = Q/T$ ，其中 Q 表示热量， T 表示温度，在台下的中国物理学大师胡刚复先生为了把这个词语解释给中国人听，用到了“商”这个概念，表明是 Q 与 T 的除法运算结果，同时其又与热量有关，因而就给“商”加上了火字旁，形成了“熵”。

如何去理解“熵”呢？在物理学领域有一种说法是“熵增”，即自然界总是朝着熵增加的方向发展。在知乎上见过有人用这个理论解释耳机线为什么会缠绕，也有人用其解释宇宙大爆炸后的现象。以耳机线缠绕为例，耳机线缠绕混乱的现象总是比不混乱时要多，这也就解释了为什么放进口袋中的耳机线取出来时总是混乱的。不混乱的耳机线是一种美好的、有序的状态，而混乱的耳机线则是一种混沌的、无序的状态，也就是说，耳机线有一种倾向从有序到无序，而无序状态的熵更大，即熵在从有序到无序的过程中增加了。

8.5.2 信息中的“熵”

从上面的例子可以知道，熵越大越不稳定，越混乱，而熵越小越稳定，越整齐。

信息学科是如何利用熵这个概念的呢？在硅被发明用来制造芯片之后，物理学和信息学的地基算是被打通了。在物理学中，稳定性高的物质熵较低，那么在信息学里，稳定性又表示什么呢？

信息学的内在是传递信息，自然地，信息传递量的大小就变成了人们关心的因素。当我们接收到别人的电子邮件，或者听到一句话的时候，时常会判断信息量大或者信息量小，而信息量大的内容对应了我们判断的确定性强，因此可以说信息量大的消息确定性强，熵也更小。通过这样的方式，我们就将物理学与信息学相互打通，也将物理学中的稳定性与信息学中的确定性相互关联起来了。

举个例子，在一个句子“I have a cat”中，重点是 I 与 cat 两个单词，因为它们能够表达这个句子的实际意义，而对于 a 这样的量词，能够传递的信息往往很有限，因此一个句子中的不同词的信息量不同。 a 这样的词，由于经常出现，可以说出现的频率比较高，出现与否并不稀奇，去除它对于我们理解这句话并没有多

少阻碍或者影响,也就是说,a传递的信息量小,故而说它给我们带来的不确定性是大的。请注意,学术界说的不确定性是单词出现的不确定性,而我们说的不确定性是这个单词对于我们心理揣度整个语境所感受到的不确定性。两者恰恰相反。

对于 cat 这样的词,由于平时出现的频率比较低,一旦出现带给我们的信息量就很大,故而我们感受到的不确定性就比较小。另一方面,如果我们能获得越多的信息,譬如把 I 与 cat 放在一起,实际上我们获得的信息量便增加了。从这个例子中我们可以看出信息量的两个属性。

减函数

信息量是信息出现频率的减函数,也就是说信息出现频率高,我们感受到不确定性大,获得的信息量就小;而出现频率低的词汇,一旦出现,带给我们的不确定性会很小,因而信息量就大。

可加性

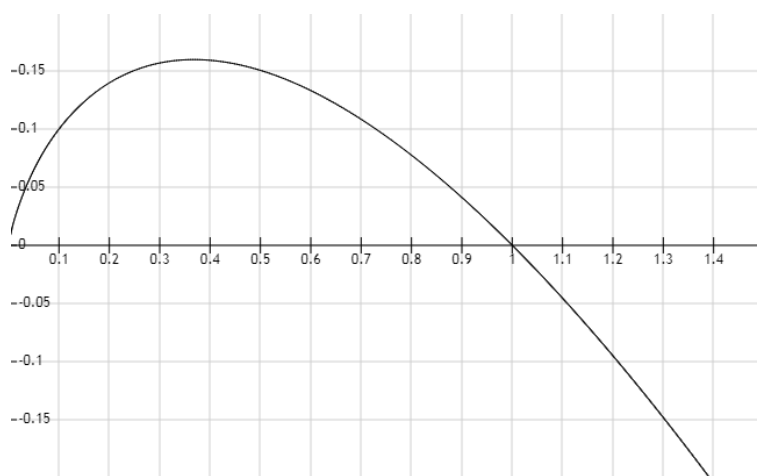
信息量具有可加性,即两个词汇的信息量是由这两个词汇各自的信息量组合而成的。我们听到一个词与听到两个词获得的信息量是不同的,通常来说,听到的词越多,信息量就越丰富,也越便于我们构建理解事物的场景。

为了找到满足上述两个属性的数学表达,我们不妨先尝试加法(即 X_1+X_2),它虽然满足第二点,但是满足不了第一点。那么如果采用 X 的倒数形式(即 $1/X$)呢?虽然满足了第一点却不满足第二点。数学总是这么神奇,使用 \log 函数能够把加法和倒数相互融合得很好,因而信息量的数学表达形式最终被确定为 $\log(1/X)$ 。而信息熵就是所有信息量的加权平均值。

$$\text{信息熵} = X_1 \log \frac{1}{X_1} + X_2 \log \frac{1}{X_2} + X_3 \log \frac{1}{X_3} + X_4 \log \frac{1}{X_4} + \dots$$

注意在上述公式中, X 表示的其实是每个单词出现的频率(或者说概率)。因而我们可以使用该公式来进一步说明为什么单词 a 会让我们不安,而单词 cat 给我们更多的确定性。

按照上述公式,每个单词其实都为产生认知提供了自己的信息。每个单词为信息熵做的贡献为 $X \log \frac{1}{X}$,变换一个形式则为 $-\log X^x$ 。我们不妨将这个公式的图像展示出来,如图 8-1 所示。

图 8-1 $-\log X^x$ 的函数图像

从上图可以看出，这个函数是一个先增后减的趋势，且拐点大致出现在 0.3 到 0.4 之间。

对于我们遇到的单词，每个单词出现的频率在整个词库中一般是很小的，因此，若单词 a 的出现概率比单词 cat 大，其为信息熵做的贡献就越大。而信息熵越大，则说明越不确定，故而我们感知到的不确定性是增加的，从而出现频率高的 a 让我们更加不安，而 cat 则给我们更多确定感。

像上面所说的那样，信息熵就是整个信息系统中的平均不确定性。对于一种极限情况，整个语料库中只有 X，因此 X 出现的频率是 1，也就是说 X 出现的概率是 100%，即任何情况下只有 X 出现，这个时候从我们的主观判断来说，状态最稳定，因为没有别的情况出现了。使用刚才的 \log 公式来进行计算，也得到熵为 0，即最小、最确定的状态，两者一致。

第9章

面向产品经理的数据挖掘

9.1	学数据挖掘，只需要高中数学	164
9.2	线性回归：人为什么没有严重两极分化	166
9.3	逻辑回归：种群增长的 S 型曲线	169
9.4	朴素贝叶斯：面相占卜工作原理	172
9.5	决策树：爱情选择背后的心理学意义	176
9.6	K-means：寻找物理学上的质心	181
9.7	层次聚类：分而治之与抱团取暖	184
9.8	DBScan：帝国崛起的定居、建国与扩张	186
9.9	关联规则挖掘：“啤酒和尿布”是个谎言	188
9.10	时间序列分析：聊聊《周易》	192
9.11	集成学习：三个臭皮匠赛过诸葛亮	195
9.12	文本挖掘：让机器读懂你	199
9.13	社交网络：隐私无处遁形	202
9.14	排序：简约而不简单的事	205
9.15	推荐系统：“今日头条”背后的秘密	208
9.16	用户画像：隐私是个“伪命题”	213
9.17	算法思想中的哲学内涵	216

从本章开始，我们终于要进入激动人心的数据挖掘领域了。不知道对前面介绍的数据清洗技术你是何感觉，统计量的知识能否消化，最难啃的骨头就要来了，我们称之为机器学习和数据挖掘。不过不要害怕，这并不是只有具备计算机背景的研究生或者坐在研究机构里的学者才可以涉足的领域，恰恰相反，这是一个来源于生活而且高于生活的领域，你只需要知道一些常识，掌握一些基本的数学知识（基本上是高中数学即可），就可以完成数据挖掘和机器学习的入门。对于那些曾经忽悠你说“这里的东西很深奥，你不必了解”的人，你可以向他们投去鄙夷的目光，然后暗暗的在心里说：“我也可以！”

9.1 学数据挖掘，只需要高中数学

本节介绍数据挖掘的时候会尽量避免使用高深的数学公式，一方面，我们承诺在给大家讲清楚数据挖掘的时候绝对不会使用超出高中数学的知识；另一方面，我们觉得大家也没有必要像算法工程师那样会编程实现。但无论如何，作为数据产品经理的你学些算法总是好的。

既然只会用到高中数学，那么很多人肯定已经觉得自己做好了。但在你热情高涨的时候，很抱歉我可能得泼一盆冷水，因为你并不一定真的掌握那些数学原理。为了更好地掌握后面的数据挖掘算法，先让我们一起温习一下基础的数学知识。

9.1.1 重温“加减乘除”

当你看到这一小节的题目时，你是不是非常震惊。难道加减乘除这么简单的道理都要复习吗？我可以负责任地说，并不是所有人都理解加减乘除。为了简便，我们只讨论加法和乘法，因为减法可以看成加上一个数的相反数，而除法则是一个数的倒数。

两个数要相加，需要满足什么前提呢？首先必须量纲相互一致。你不可以把一头牛的重量和一个人的身高相加，也不能够把两个采取不同单位进行高度度量（一个是 1.72 米，另一个是 181 厘米）的结果相加。当满足了这一条之后，就可

以说两个数进行加法运算没有障碍了。在这个基础上，两个数相加满足的是交换率和结合律。

对于乘法来说，它是多个相同数做加法的重复，因而除了满足加法的交换率与结合律外，还有着自己独特的定律，即分配率。

而多个乘法重复就形成了幂级数，幂的运算比乘法运算少了一个交换率。这就是我们重新认识的加减乘除。

9.1.2 重温“比值”

当说到比值，我们不妨先回顾一道小学数学题： $1/2$ 比 $1/3$ 多多少？你的答案是 $1/6$ 吗？恭喜你，答错了。答案应该是 $1/2$ 。因为一个分数（比值 A）比另外一个分数（比值 B）多多少，这个问的实际上是 A 比 B 多的部分占 B 的多少。如果你怀疑我是在文字上故意没有讲清楚，那么不妨回到小学课本上去找找原本的定义，并非我诓你，真的是定义使然。

在后面介绍数据挖掘算法的时候，我们时常会遇到分数，例如比、比值、比例、比率、频率、概率，如果不区分特定的场景，可以近似认为它们表达的同一个概念。

9.1.3 重温“函数”

函数就是映射。映射就好比海上升起的一轮明月，天上有个月亮，海中也有个月亮，真实的月亮倒映在水中，相当于一个物体映出了另一个物体。真实的月亮有阴晴圆缺，水中的月亮也会跟着变化，这就是函数，他们不仅仅是倒映关系，还是一种相互依存的对应变化关系。

9.1.4 重温“符号”

对于符号，我不会介绍深奥的微积分知识。这里我只说两个高中时候就学过的符号，一个是自然对数 e ，另外一个自然对数的相反函数 \log 。

我们都知道 e 是一个常数，和圆周率 π 一样，可是 e 究竟是怎么来的呢？ e 最初源于经济学。早年在欧洲地区，人们虽然生活在封建主义时期，但是资本主义萌芽已经兴起，人们往往会进行贸易往来，并形成借贷。借贷就必然会遇到利息，自然对数 e 和收利息的方式有关。

假设一个人借出 100 元，共借一年，利息收 100%，即到一年后收回利息与本金共 200 元，那么他赚了 100 元。如果分一年两次收，每次收的利息仅仅是一年利息的一半（即 50%），那么可以收到 $100+100 \times 50\%+100 \times 50\%=225$ ，赚了 125 元。如果一年分 4 次收，每个季度利息是全年利息的 1/4，那么可以收到 $100 \times (1+25\%)^4=244.14$ 元。如果改成每个月收一次，每次利息变为 1/12。如果进一步按照这样的规则细分呢？每天收一次？或者每小时收一次？而利息进一步摊薄，结果会如何？

看起来金额是越来越大，但是不会无限制增大，其会收敛到一个值，最后能够收到的钱，其实就是 $100 \times e$ 。这就是 e 的来历。而 \log 则不用多说，它和 e 是一对孪生兄弟，我们一般遇到最多的就是以 e 为底的对数，称之为 \ln ，后文中如果不特别指出， \log 就是指 \ln 。

e 与 \log 的出现改变了加减乘除。原来加和乘是不能够相互转换的，但是通过 e 与 \log 就可以把加法和乘法进行互换，譬如 $e^2 \times e^3 = e^{2+3}$ 。

9.2 线性回归：人为什么没有严重两极分化

线性回归由两个词组成，一个是线性，一个是回归。线性很容易理解，就是我们曾经在高中的时候学过的坐标系里面的直线。另外一个词“回归”要怎么解释呢？这就不得不回到线性回归的历史源头去追溯了。

9.2.1 优生学趣闻

有一个名为弗朗西斯·高尔顿（Francis Galton）的人，他是博物学家达尔文的表兄（另一说法为表弟），也是一名生理学家。由于其与达尔文关系密切，因而深受进化论思想的影响，并且把进化的思想带入了自己的研究中。在研究个体差

异的过程中，他着力从遗传学的角度找原因，从而开创了优生学。

对于优生学，大部分中国人并不陌生。优生学旨在通过预防的手段，减少不良个体（例如具有先天缺陷或者严重遗传病）出生，从这个角度来看，优生学是有积极意义的，提高了出生婴儿日后的生活质量。但正如所有的事物都是双刃剑一样，优生学也在二战期间被纳粹拿来鼓吹雅利安人的优越性，从而为其残酷的种族灭绝政策掩人耳目。从这个角度来看，优生学也有消极的影响。

Galton 老先生的本意自然并非如此，在 1855 年的时候，Galton 研究了 1078 对父子的身高数据，这个数据在当时应该算是比较大的样本数据了。通过研究发现，这 1078 对父子的身高满足一个公式，也就是 $Y=0.8567+0.516 \times X$ ，其中 X 是父亲的身高，而 Y 是儿子的身高。这个公式是我们都知道的二元一次方程，要是画在坐标系里面就是一条直线，因而称之为线性。

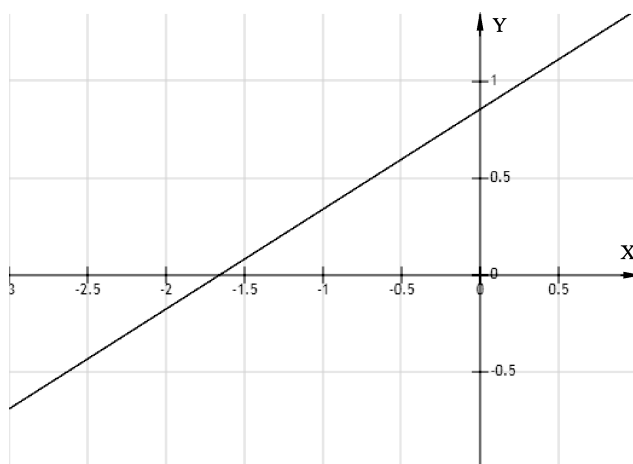
回归是怎么一回事呢？Galton 通过数据发现，尽管大体来说高个子的父亲有高个子的儿子，个子矮小的父亲有个子矮小的儿子，但是情况并非总是这样。当父亲特别高或者父亲特别矮的时候，往往孩子会相反地比其父亲更矮一些或者更高一些，不太会按照更加高大和更加矮小的方向进行发育。这种现象仿佛说明人的身高数据有一个中心，人们的身高总是围绕着这个中心，当太过于偏离中心的时候，就会有被拉回中心的趋势，因而称之为回归。

至此，“线性回归”这个名词就完全得出且实至名归了。

从生物学的角度来看，人的各种属性必然具有回归性。就身高而言，如果高者更高，矮者更矮，那么人类就会在数百万年的发展过程中分化为两个或者更多的群体，而事实上并没有出现这样的情况，巨人症或者侏儒症仅仅是少数特例罢了。所以，不要忘记 Galton 为我们揭示的回归现象。

9.2.2 空间中的直线

上述讨论的父亲与儿子身高关系的表达式 $Y=0.8567+0.516X$ 显示的是一个 X 与一个 Y 的关系，我们可以把其数学函数的图像画出来，如图 9-1 所示。

图 9-1 $Y=0.8567+0.516X$ 的函数图像

这便是线性回归，从图中可以看出，这样的函数是在一个二维的平面上呈现的，一方面意味着整个函数只有两个变量，另一方面表明影响儿子身高的因素仅仅只有父亲身高这么一个因素。但是如果影响身高的因素有很多呢？我们是不是也可以用线性的公式进行表示呢？

例如影响儿子身高的因素除了父亲的身高之外，还可能有母亲的身高、营养状况（可以由经济状况表示）、三代以内有无侏儒症等因素。针对这样的情况， Y 没有变化，但是 X 从一个因素变成了四个因素。某种程度上来说我们仍可以使用线性回归的方式来找出儿子身高与上述因素之间的线性关系，不同的仅仅是 X 变成多个，即 X_1 、 X_2 、 X_3 、 X_4 。我们称 X 为自变量，也就是自己会变化的量，而 Y 称作因变量，也就是因为 X 变化而引起变化的量。

在上述情况下，图像已经无法在一个二维平面中展示了，可以扩展到三维空间，甚至五维空间。在多维坐标中，其中一个变量就是儿子的身高，另外的变量是其他影响身高的因素。

这就是线性回归，它能够用来探索多个变量与另一个变量之间的线性关系。所以下一次如果在实际的数据中想找出若干个变量之间的关系，首先要分清楚谁是因谁是果，然后按照<自变量，因变量>的格式整理出数据，这样就可以使用线性回归找到数据之间的关系了。至于用什么编程语言实践，各位数据产品经理们可以向数据分析师们请教，或者通过自学 Weka 等软件来实现。

9.3 逻辑回归：种群增长的 S 型曲线

说完线性回归，我们要说逻辑回归。我们已经明白了什么是回归，那么什么是逻辑呢？这个名词似乎和线性完全没有关联，也不好理解，难道需要使用逻辑推理不成？其实不然，逻辑回归是汉语音译罢了，其真正的英文是 logistic，英文发音是“逻辑斯谛”，所以称 logistic regression 为逻辑回归。

9.3.1 种群的增长曲线

那么 logistic 究竟是怎么来的呢？要想弄清楚，就要讲到另外一个科学家，皮埃尔·弗朗索瓦·韦吕勒（Pierre François Verhulst）。

韦吕勒出生于 19 世纪初（1804 年）比利时的布鲁塞尔，19 世纪 30 年代初，比利时独立，国家开始迅速发展与扩张，从自由资本主义向帝国主义过渡。韦吕勒作为社会学家和数学家开始崭露头角。1941 年，他当选为比利时皇家科学院院士，并在其去世的前一年当选为科学院主席。期间，他通过数据分析给出了影响人口的因素，并且估算比利时人口上限为 9 400 000。一个半世纪过去了，根据世界银行公布的数据（<https://data.worldbank.org/country/BE>），比利时的人口在 2000 年左右刚刚迈过 1000 万，这与韦吕勒的估计相差无几。

韦吕勒研究人口数量的思路可以推广到生物领域，并找到一定的理论源泉。在生物学中有一对概念，称为“J 型曲线”与“S 型曲线”。

J 型曲线指的是物种数量的增长随着时间的变化关系呈现类似于“J”的指数型增长效应。这是由于在空间充裕、气候适宜且没有天敌的情况下，种群的数量每年会以一定的倍数增长。

与之相对应的 S 型曲线，则描述了一种有“天花板”的增长情况。这是由于自然界的资源和空间总是有限的，当种群的物种密度增大时，种群内便出现了斗争。物种越来越多，竞争越发激烈，且以该种群为食物的物种数量也会相应增加。综合因素造成了该物种出生率降低，死亡率增高。当死亡率升高到与出生率相等的时候，种群的增长就会停止，从而触达“天花板”。韦吕勒对比利时人口的上限统计便是 S 型曲线的一种情况。

他在 1844 年到 1845 年期间研究物种数量，特别是观察种群中人口数量的增长时，发现了一个规律，即刚开始的时候人口增长缓慢，然后越来越迅速，这个过程称为增长期。当人口或者物种增加到一定数量的时候就出现了社会资源的瓶颈，大家开始竞争，于是开始有个体死亡，种群数量增长开始变缓，直至饱和。整个过程如果在图上画出来非常类似字母“S”，如图 9-2 所示，故而称之为 sigmoid 曲线（意为 S 形状的曲线）。

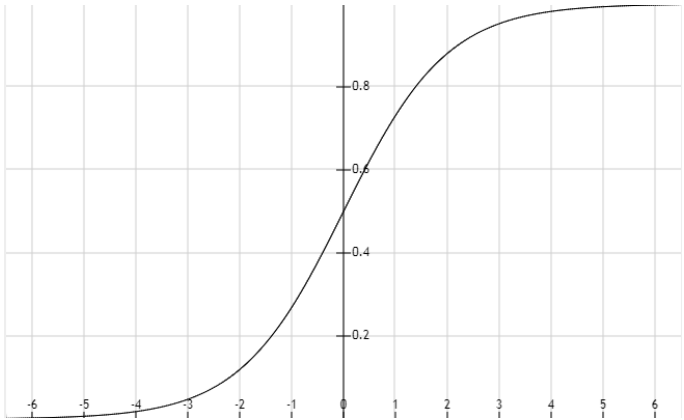


图 9-2 sigmoid 函数图像

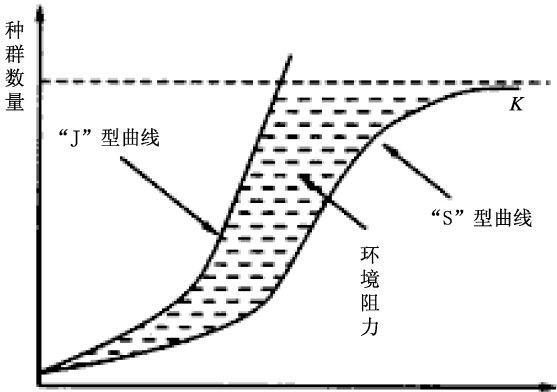


图 9-3 生物种群数量中的 S 型曲线与 J 型曲线

9.3.2 S 型曲线的秘密

我们常常使用 S 曲线来形容一个女性傲人的身材，如果把整个种群看作一个整体，那么社会生物女性的 S 曲线究竟隐藏着什么样的秘密呢？让我们一点点来揭示。

这样的曲线有两种不同的名称，意思是一样的。一种叫法是 sigmoid 曲线，而另外一种则是 logistic 曲线。前者是按照其函数图像的“长相”来命名的，而后者则是根据函数图像的数学形式进行命名的。

S 型曲线的数学形式是 $y = \frac{1}{1+e^{-x}}$ 。假设对这个函数进行求反函数的操作，会得到什么呢？我们会得到一个 log 函数，具体形式为 $x = \log\left(\frac{y}{1-y}\right)$ 。

英文中表示奇数、偶数分别使用的是 odd 和 even 这两个单词。odd 可以看成来源于单词 add。按照元音互换理论，odd 与 add 有相似的含义。众所周知，add 表示加，而在原来已有的事物上加上新的东西必然和原有的事物不能很好融入，有些奇怪和特立独行，故而 odd 表示奇特、古怪的意思，进一步演变成落单的含义，也就是现在的“奇数”（单数）。

英文中在 odd 这个词后加上一个后缀 s 变成 odds，表示几率。几率和我们了解的概率略有不同。如果把 Y 看成事物发生的概率，表示的是事物发生的次数占总次数的比例，于是 1-Y 就是事物不发生的概率，而 $Y/(1-Y)$ 就是事情的几率， $\log(\text{几率})$ 就称之为对数几率。

logistic 以 log 开头，表明的就是这种对数几率。如果把对数几率看成是一个整体因变量，那么对数几率就可以表示为 x 的一个线性函数。这样，逻辑回归与上个部分讲到的线性回归便联系到一起了。

书归正传，回到 S 型曲线的图形上，我们可以观察到逻辑回归中的因变量 Y 具备这样的特性，即相较于线性函数来说，在 0.5 的分界处，其取值开始极速增大或者极速衰减，从而使得其快速偏向于 1 或者 0。这样的性质有一个好处，我们可以用它来表明 1 和 0 两个完全不同的类别，不会使得中间的过渡阶段过于模糊而难以分别。打个不恰当的比方，逻辑回归的 S 型曲线是一个较为果断的人，

它对于非黑即白（非 0 即 1）的判别比线性回归要更加果断，不会犹豫和模棱两可太长时间。所以我们可以使用逻辑回归来进行两类事件的区分。

如果将线性回归和逻辑回归放在一起比较，我们大致可以得出如下结论。

目的不同

线性回归是为了研究因素之间的线性影响关系，本质上是拟合；而逻辑回归则用来进行两类事物的区分，本质上是分类。

函数形式不同

逻辑回归在线性回归的自变量组合后，还增加了一层 sigmoid 函数；如果把对数几率当成一个完整的因变量，两者的数学形式是一样的。

因变量 Y 值域不同

值域就是取值范围的意思，在线性回归中，Y 可以从负无穷到正无穷，而在逻辑回归中则只能在 0~1 之间徘徊，这也是由 sigmoid 函数属性决定的。

多嘴说一句，其实两者在计算和推导的逻辑上也是不同的，线性回归要实现平方和最小的目标，而逻辑回归则要实现 log 损失最小的目标，类似的 SVM 则要实现 Hinge 损失最小的目标。如果未来有机会，我会专门针对机器学习过程中的目标函数优化问题做出详细诠释，如果觉得本段内容过分深奥，我建议产品经理从故事入手了解来历，在日后的深入学习中需要辅助理解时再回看这些内容即可。

9.4 朴素贝叶斯：面相占卜工作原理

见识了回归二兄弟，我们要来严肃地说一说分类这件事情了。我们会习惯性地看到的事物以及见到的人进行分类，因为分类是对未知世界的一种简化。简化和抽象是人们认识世界的两种最基本的模式，通过简化，信息可以以最快的速度传递，通过抽象，则可以以最深的程度认识信息。

之前介绍的逻辑回归方法是一种分类方法，因为它可以将事物分成两类，也就是“是”与“非”二类，这将是很多分类的基础。在一个时事出英雄的时代，不可能仅仅只有一个人提出科学观点和科学方法，托马斯·贝叶斯（Thomas Bayes）就是和同时代众多科学豪杰并举的一位人杰，其在统计推断领域做出的贡献巨大，

开创了属于自己的理论“贝叶斯理论”，围绕这一理论进行研究的人被称为“贝叶斯学派”。毫不夸张地说，他获得了比同时代科学家更高的知名度、美誉度与传颂度。下面就让我们进入贝叶斯的分类世界。

我们判断一个事物应该属于 A 或者 B 其中的哪一类，本质上是判断这个事物属于 A 或者 B 的概率哪个更大？这样的方式在学术上称为“软分类”，也就是承认这个世界并非非黑即白，而是存在灰色的中间地带。例如在改革开放早期关于姓资还是姓社的讨论，就是一个二分类的问题。许多人一定要简单地把社会定型下来，这就是一种二元思维，学术上称为“硬分类”。历史上的 1933 年，发生在美国的罗斯福新政支持了资本主义也有计划经济的结论，从而进一步巩固了软分类的理论。

9.4.1 外貌协会与街头看相

在国内的年轻人中，特别是年轻女性中，流行着这样的一个词汇——外貌协会（Good Looks Club）。这个词特指一群奉行外貌至上，颜值为先的人，而脸蛋则是他们在外貌中最为关注的身体特征之一，他们具备了看一眼脸蛋就知道这个人的习性与品格的能力，甚是神奇。这不禁让我联想到街边看面相的江湖术士，他们也可以根据一个人的面相来预测旦夕祸福，有异曲同工之妙。

这背后的工作原理和贝叶斯有关。为了简便讨论，我们不妨把人的处境分为福与祸，占卜的剖析过程可以分为下面几种。

连蒙带猜

俗话说“福兮祸所依，祸兮福所倚”，也就是说有福就一定存在一个与之依存的祸。这样说来，福祸发生应该是一半一半的。可以据此来简单预测。

社会常识

虽然福祸依存相互交错出现，但是从整个社会中看来，幸福与快乐总是大多数，我们并没有看到与福等量齐观的祸。可以依比例来预测旦夕祸福。

阅历使然

术士们阅人之时从客户身上提取到个人特征，除此之外还可能包括衣着、谈

吐等，进行预测。

此处为了简便，只讨论某种面相对于预测吉凶的影响。我们不妨假设术士曾阅过 1000 个人，其中 100 个祸，900 个福，大体看来与社会常识并无差别。但当观察面相的时候，他发现祸中有 40 个存在相同的面相，而福中这样面相的数量是 130。问题来了，他究竟该针对拥有该面相的客户如何预测其福祸呢？

术士的内心过程可能是这样的：我遇到的 1000 个人中，出现了 100 个祸，所以出现祸的概率是 $100/1000$ ，而在其中有这个面相的比例则是 $40/100$ ，两个相乘起来就是 $40/1000$ 。同样，也可以按照这个方法针对福与福中该面相的人计算一个概率，于是得到 $130/1000$ 。两个概率相加就是所有群体中有这个面相人的比例，即 $170/1000$ ，而在这个概率之中，福贡献的显然要多于祸贡献的（比例是 $130:40$ ），所以应该判定有这样面相之人属于有福之人。

在上面提到的几种预测方案中，方案二属于频率学派的解法，他们专注于整体的比例，通过不停抽样调查获得事物的真实情况；方案三就是贝叶斯学派的做法，他们不仅仅专注于整体的已知比例，还将一些观察到的细节和经验加入进来。

在贝叶斯学派看来，从面相判断一个人经历的事是福是祸，有两个步骤：首先得知道福和祸的各自比例；其次是知道福和祸中该面相的比例，然后相乘，哪个结果大就选哪个。其中第一步是频率学派已经具有的，而第二步则是其缺少的，我们一般称第一个步骤为常识，也就是先入为主的经验，学术界称为先验；对于第二个步骤，我们需要从数据中去发现，其实是添加了面相这样的信息加以修正，对我们判断是福是祸的原先经验起到了修正的作用，故而称之为后来修正过的经验，学术界称之为后验。如果后来添加的特征信息（如本例中的面相）变得更多，更细，便会得到更为精准的后验知识。

由此可知，贝叶斯学派和频率学派本身的出发点就不同，频率学派认为这个世界上万事万物的各类占比都是一定的。而贝叶斯学派则使用了额外信息，认为世界在不断变化，人们对事物的看法应该是不断修正的。

9.4.2 无处不在的贝叶斯

上面用江湖术士看面相占卜福祸的例子来介绍贝叶斯理论，算是一种戏谑，

想必术士们未必懂得概率与贝叶斯。尽管如此，我们不能否认贝叶斯理论的用处。事实上，它从被发明的那一天起，就注定要被广泛使用。

垃圾邮件的屏蔽

你只要使用邮箱进行办公，就一定和垃圾邮件打过交道。收到垃圾邮件有可能是因为你曾经注册了一些账号，从而造成长期订阅邮件的侵扰；或是由于信息泄露，而收到定向广告的宣传。无论如何，垃圾邮件夹杂在工作邮件和一些非常重要的邮件中，会影响工作，甚至造成误删某个重要邮件的损失。

现在大多数邮箱都提供了一种功能，叫作垃圾邮件屏蔽。这样的功能使得邮箱系统能够对你邮箱收到的邮件进行自动分类，从而把它认为可能是垃圾邮件的一部分邮件屏蔽掉。而在产品中的具体做法便是把这群被认为是垃圾的邮件挪到一个专门的文件夹中，使收件箱清爽，只展示那些重要的邮件。

屏蔽垃圾邮件是怎么做到的呢？我们通常会通过一些关键词，例如优惠、爆炸价、地铁旺铺等，来识别我们认为是垃圾邮件。但事实并非总是这样，如果我是一名地产代表，邮件中出现地铁旺铺就是一个正常的现象，所以不能一概而论。对于系统来说，它需要有能力识别那些有极大可能属于垃圾的邮件。这个过程事实上是在垃圾邮件的特征（如关键词、发送人邮箱、正文链接地址等）和是否为垃圾邮件之间建立因果联系。而系统需要在用户收到邮件的时候就反过来根据这些特征推断究竟是否属于垃圾邮件。相似地，系统也在为邮件看面相，特征便是邮件中提取的各种属性。

检测结果假阳性

在医学检测中，通常使用阴性表示正常，而使用阳性表示异常。但结果往往也会由于某些特殊情况而产生变化，从而出现两种差错：第一种差错是，原本是阳性却没有检测出来；第二种便是，原本是阴性却被检测出是阳性。后一种我们称为假阳性。假阳性对于患者的伤害是比较大的，这等于误诊宣判一个健康的人有病。

曾有人建议在新婚检查的时候进行血液检查，对是否患有某种疾病进行检测。但该计划提出后便遭到专家的质疑，并最终搁浅。专家之所以质疑，原因在于，在一个发病率为千分之一的国家，如果试剂检测出阴性的准确率为 99%，而检测

出阳性的准确率为 95%，那么当试剂在一次检测中发现了阳性，其真正的患病概率却不足 9%。如果因为这么小的概率就在婚前拆散了一对男女，定会引发巨大的社会问题。

在这个过程中，专家实际上是在是否患病与阳性与否之间建立因果联系，然后再反过来使用阳性与否的结果反推受试者是否患病。在这个过程中专家又一次扮演了术士的角色，这次的面相是试剂的阳性结果。

无法逃脱的罪责

美国历史上有这样一幕。1981 年 3 月 30 日，一名大学生 Hinckley 企图对里根总统行刺，并在行刺过程中伤及数人。在后来的审判中，其代理律师引用了 CAT 扫描的结论证明 Hinckley 有脑萎缩，并借此向法庭申诉其可能患有精神分裂症，从而免于罪责。

但陪审团和法官使用了贝叶斯的理论，调查了美国当地的精神病发病率，这一数字为 1.5%，并计算出由于精神疾病而得脑萎缩的概率也只有不到 20%。因此未能以此为理由让 Hinckley 开罪。在这一过程中，法官找出了精神疾病与脑萎缩之间的因果关系，并且由结果（脑萎缩）反推可能患有精神疾病的概率。法官扮演了术士的角色，其所看的面相便是脑萎缩病人的特征。

以上三个例子分别介绍了贝叶斯理论在不同场景下的作用。实际上其应用场景不仅限于此，但从中我们能够找到一个共性，即使用贝叶斯理论的时候，往往需要找到一对因果关系，并且详细观察特征（面相），然后基于结果反推原因是什么。

很多场景下，我们会有一些误解，这是因为我们把因果倒置了。例如误以为阳性就是患病，殊不知因为患病才会有阳性。这种逻辑谬误好在有贝叶斯理论可以帮助修正。这就是贝叶斯理论，一个依靠信息修正认识的方法。

9.5 决策树：爱情选择背后的心理学意义

我们沿着数据挖掘的康庄大道继续前进，在来时的路上，我们谈到线性回归起源于生物统计学，逻辑回归最早应用于物种社会统计学，贝叶斯理论则是重新

认识世界的一种哲学价值观。那么本节中要聊的决策树又有着怎么样的渊源呢？

这里得提到一位有名的澳洲科学家罗斯·昆兰（Ross Quinlan），昆兰在 1965 年从悉尼大学获得物理和计算科学（Physics and Computing）学士学位后，又前往华盛顿大学攻读博士，后又在悉尼大学和悉尼科技大学任教。

2011 年，昆兰更是名噪一时，他获得了数据挖掘领域的最高成就奖项“KDD 创新奖”。尽管现在昆兰已经被计算机领域的人士熟知为一名伟大的机器学习专家，但是很少有人知道他是在研究心理学的时候发现了令他在计算机领域名垂青史的算法。

昆兰最早是研究人们在认知过程中如何进行概念学习的，也就是人们在学习一个概念的时候是如何进行思考和心理活动的。他发现了人们经常会采用一种“分而治之”的思考方法和策略，其过程就是不停地进行判断，然后最终抵达问题的根源与尽头。这种方式有点像不停地反思和自问自答，最终将一个事物和概念建立联系。

就是在研究这样的概念学习的过程中，昆兰发明了使他名震数据挖掘圈的算法“决策树”。决策树究竟是什么？且听我慢慢分解。

9.5.1 爱情选择条件多

决策树的名字中既包含算法目的和使用场景，也包含算法直观呈现，这在数据挖掘算法中十分难得。其中“决策”表明了这个算法可以用来进行决定和选择。最简单的选择是 YES 或者 NO，这在学术上称为二分类问题，我们在项目流程图中使用菱形图来进行条件判别也是一个例证。另外，“树”这个字透露了算法工作的原理，只不过这个树是一个树根在上树冠在下的倒置的形状。从树根开始，我们沿着树干或者枝杈直达树叶，每每遇到树枝分叉的时候，我们就停下来选择前进的方向。这个时候就要使用一套准则来决定究竟向哪个方向进发。

假设我是一只渺小的蚂蚁，想要找到某棵树的某片叶子上存在的一份宝藏。我需要从树根开始走到树叶，寻找那份宝藏。但我并不会盲目寻找，在从根部出发前会有很多的寻宝图，这些图中的信息可以在我每次遇到岔路的时候告诉我该往哪里去，并最终指引我到达一片叶子。然而很可惜只有一张寻宝图是有效的，

而且每次到达叶子后如果发现没有宝藏就得回到根节点重新拿寻宝图。在这个过程中，寻宝图是决策树中用于决策的特征和信息，而宝藏则是最终的结果，寻访的过程就是决策树算法运行的过程。

让我们再来举个例子，譬如现在有一个姑娘打算找对象，她会怎么对她的恋爱对象进行选择呢？对应上述寻找宝藏的过程，寻宝图中罗列条件是必须的，如身高、体重、收入以及职业等。依据这些条件进行爱情选择的过程就是寻宝的过程，也就是姑娘进行决策的过程。

这个过程大致会是这样：她会把这些因素按照自己关心的重要程度由深到浅进行排列。接下来的每一次相亲，她就会使用这样的条件去判定一个人究竟适不适合做未来的伴侣。规则的形式可以是“首先身高必须比我高；在身高条件达到的情况下，收入得能买得起房；如果上面两项都满足的话，最好家乡离我家近一些，以及工作别太忙有时间陪我”。这就是进行爱情选择时的决策树。

不妨抽象一下刚才这位姑娘的决策过程，首先是列举特征，然后是按照重要程度排好序，并且设定一系列规则，把重要程度深的放在决策树的最上面优先判断，程度次要的依次向下。

不过大多数人会有所纠结，这么多的条件，难以抉择哪个是最为重要的，那么决策树有没有给我们提供一些解决这种问题的方案呢？如果有，具体是怎么解决的呢？

9.5.2 不纠结的小技巧

上面针对各种条件排序的问题，可以等价于判别条件和特征的重要性。大致有三种方法可以用于条件重要性的评判，其基本原理是一样的，即让条件和特征相互竞争，以输赢定轻重。这三种方法在特征间相互比较的时候，采用的评判标准是不同的。

少错即是对

错与对本身就是一组正反义词，判断特征重要性的一个依据就是看根据这个特征做出的判别是否能够指引我们达到最终的目的。对于爱情与婚姻，衡量的标准姑且设定为婚后的忠贞与专一。仍以姑娘选择对象这件事为例，例如使用身高

这个特征作为筛选条件，判断个高的婚后比较忠诚，若按照这个规则进行筛选，会发现 10 个人中有 5 个都判断错了。对于这种使用身高特征进行决策树的构造来说，错误率就是 50%。若换而使用职业这个特征判定，譬如认为程序员更为专一，依据此进行决策树判别，发现 10 个人中仅有 2 个人被判断失误。对比上述身高和职业两个特征，我们就可以说利用职业这个特征判断忠诚度比利用身高判断更准确，因而职业这个特征就胜出了，表明比较重要。

熵小就是强

使用错误率进行判别是一种最为简单的方式。昆兰发明了另外一种用来判别特征重要性的标准，就是熵。我们曾经提到过，熵越大越不稳定。那么对于决策过程来说，什么样的特征才是重要的呢？很显然，如果一个特征能够像一把快刀一样将有待判断的对象一劈两半，并且分开的两个部分都各自只属于一种类别，那么这样的特征就是我们最想要的，对我们来说也是最重要的。例如，我们发现将男女双方家乡距离远近这个特征作为划分标准对判断对象进行划分后，其中一个部分的人都是专一的，而另一个部分的人都不专一，这样我们就可以仅仅依据这一个特征找到让自己心安的未来真爱了。

让我们通过熵的视角来观察划分前以及划分后的这两个部分。在划分前，整体中存在两种类别，这个整体有一定的熵值。当划分完成后，两个部分中各自都仅有一个类别，这个时候熵为 0，是最小的。其他的任何划分都不可能获得比熵为 0 更好的结果。于是我们可以下这样的论断：好的特征就是，使用这个特征划分之后，熵要尽可能小。熵变得越小，熵减少的程度越多，这个特征就越重要。依据这个思想设计的决策树算法有两个，一个叫作 ID3，另一个叫作 C4.5，区别就在于前者适合离散型数据使用，而后者适合连续型数据使用。

稳定确实好

既然决策树这么好用，为其设计“不纠结”条件筛选的人肯定不止一个，里奥·布莱曼（Leo Breiman）就是其一。布莱曼在 1984 年与另外几人一同提出了一种决策树算法：分类与回归树算法（Classification and Regression Trees, CART）。2005 年，布莱曼老先生去世，但他留给我们数据挖掘领域的这笔宝贵财富却永远不会被忘记。

简单来说，CART 算法就是将决策树中用于判断特征重要的方式由 ID3 的信

息熵变成了基尼（Gini）系数。基尼系数最早是由美国经济学家阿尔伯特·赫希曼（Albert Otto Hirschman）提出的，用于衡量社会收入分配是否公平。

以图 9-4 为例测算基尼系数，先将社会上所有的人（或采样的人）按照收入由低到高排序，然后按照累计人口来计算累计收入。例如有五个人，月薪分别是 1000、2000、3000、4000、5000，累计收入就是 1000、3000、6000、10000、15000。如果按照这样的数据在二维坐标上画出图像，就会得到和图中向下弯曲的线，学术上称为洛伦兹曲线。但如果现在情况变成五个人的月薪都是 3000，那么累计收入就变成了 3000、6000、9000、12000、15000。在这个情况下作图，就会得到图中的直线。

试问这两种收入情况，哪一种更加公平？毋庸置疑是第二种。于是我们可以得到一个结论：洛伦兹曲线越接近于收入分配绝对平均的直线，社会就越公平。基尼系数由图上的两块面积定义，即 $A/(A+B)$ ，这样在绝对公平的时候，基尼系数最小，为 0；而当社会极度不公平，仅有一个人占有所有财富的时候，基尼系数最大，为 1。基尼系数越小，社会越稳定。

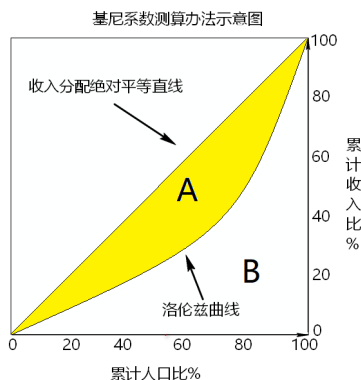


图 9-4 基尼系数计算示意图

基尼系数和我们的决策树又有何联系呢？一个好的特征应该是可以将所有待判别对象明确分开的。布莱曼就是从基尼系数中得到启发，提出了基尼指数的计算方法。在使用某特征进行划分后，会计算得到两个部分的基尼指数，基尼指数越小，则说明这种划分就越靠谱、越稳定（类比社会稳定）。特征的重要性取决于哪个特征能够最大程度减小基尼系数，使基尼系数最小的那个特征划分便是对我

们来说最重要的特征。

至此，我们已经掌握了三种用来判断条件重要性的方法。这三种方法虽然要求有数学功底，可一旦掌握，便再也不用纠结，岂不美哉？

9.6 K-means：寻找物理学上的质心

文前提到的逻辑回归、贝叶斯以及决策树等方法虽然各不相同，但殊途同归，大家都是奔着一个目标去的，那就是对事物进行分类。分类两个字看似简单，其实奥妙无穷。分类蕴含了一个既定的事实，即这个类别已经存在，而每个类别的存在又是以该类别包含一些实际的例子或者样本为前提的。

与之相关联的还有另外一个概念，称为聚类。聚类和分类虽然仅仅一字之差，但意义却截然不同。首先，聚类不一定需要这个类别已经存在。聚类过程本质上是寻找相似事物的过程。若找到，则与之合并；若没找到，则自成一类。相似的事物只有聚集在一起的时候，才能成为一个类，在此之前并没有规定好聚在一起的是什么。其次，正如刚才所说，分类中的每个类内已经有了代表成员，而聚类则看成在招揽代表成员。我们找不到这个特定类的标准，因为它随时可能会变换。这个类中的成员在不停地增加，不停地变动，因而能够代表这个类的成员也不固定，故而我们说聚成的类没有一个特定的严格标准。

按照上面的说法，一方面，聚类似乎是分类的前序过程，分类过程中每一个类的成员可以通过聚类得到；另一方面，分类又是聚类的核心诉求，聚类的过程实际上也蕴含着分类的思想。从本节开始，我们将介绍若干聚类算法，先从 Kmeans 开始吧。

9.6.1 向中心看齐

先从名字谈起吧，根据梅林韦氏词典，mean 的解释中有一项是 occupying a middle position（意思为占据中间的位置），翻译过来就是平均数的意思。K 则和最终需要聚成类别的数量相关，K 可以取 1、2、3 之类的自然整数，表示最终我们需要聚成多少个类别。

那 K-means 的过程是怎么样的呢？让我们用社会中建立社团的例子来打个形象的比方。在社会中，会有各色的社团，有的社团是功能型的，例如红十字会、保护动物协会等；有的社团是兴趣聚集型的，例如某明星粉丝团、谷歌粉丝团（约十年前有一个报导谷歌奥秘的网站叫作谷奥，可以算是较早的自媒体社区，我有幸能够成为其编辑之一）等。

无论是功能型还是兴趣型社团，都是从小变大的。过程中必不可少的是核心人物的作用，这样的人我们通常称之为意见领袖或者中心人物。“英雄造时势，时势需英雄”，社会中若即将成立 K 个社团，则必先要有 K 个中心人物作为领导者出现。于是我们先要从芸芸众生中找出 K 个人作为即将成立的 K 个社团的中心人物。这个时候社团实力相对较为薄弱，为了吸纳新的成员，扩大影响力，社团需要对外宣传与介绍自己。这个过程就是社团不停树立自己品牌形象的过程。对于一个未曾属于任何社团的新人来说，他选择加入某个社团的决策过程就是进行聚类的过程。在决策过程中，他会综合考虑这 K 个社团的特色以及这 K 位中心人物的特点，并最终选择那个与自己情投意合的社团和高度认同的领袖。这个认同的过程在数学上则表现为相似度，我们通常所说的“不是一家人，不进一家门”便是这样的道理。

当该新人选定了某社团后，整个社会中发生的变化有，某社团增加了一名成员（中心领袖与该新人），其余的社团仍为中心领袖一人，社会中未参加社团的人数减少了一位（即该位新人）。随着社会中每一位新人加入了与之相匹配的社团，最终社会中再也没有新人了，而社团也因有更多的人加入而羽翼丰满。整个过程结束后，没有发生的变化是社会社团的数目，仍为 K 个。

在最开始的时候，我们并不知道选出的这 K 个人即将成立的社团叫什么？仿佛对于社团组织成立的定位并不清晰。但随着不停有新成员加入，社团不断调整自己的方针与策略、主张与口号，渐渐地拥有了一个明确的主题和目标。对于一个外人来说，这个社团是做什么的，只要看看这个社团中的成员就知道了。于是在社团吸纳成员（聚类过程）完结后，每个社团都可以取一个响当当的名字（类别冠名）。至此，K-means 的聚类过程结束，我们得到了 K 个社团（类别）。

在物理学中有一个概念叫作质心，其意义是质量的中心，表明物体的质量集中于这个点上。这个点可以选作整个物体的代表。在经过算法聚类而成的每个类

别中，并没有一个真正的类似于真实社团组织中领导者的角色。每个聚类后的类别，甚至是聚类过程中动态生成的类别，很难说清楚其中心是谁。可以认为，这个类别是由所有类中成员共同决定的。如此一来，每个类别中都可以有一个虚拟的领袖，这个领袖是所有成员气质和特点的平均，他代表了这个类别。如此一来物理学中的质心便有了现实的意义。我们可以说，它是社团中的领袖，也是样本聚类后的类别中心。

这被称为社会物理学。以 K-means 聚类为例，它不过是把物理学上的概念抽象之后进行算法化，从而形成了最早期的聚类算法。在实际生活中，只要我们善于观察，就可以把一个领域的知识迁移到另外一个领域进行创新，这种跨界的创新在当下显得尤为珍贵和重要。这是跨学科的意义，也是促使我们博学的动力。

9.6.2 站错队的后果

K-means 可以用于解释社团与组织的形成，但正如社团会有危机一样，K-means 也有自己天生的短板。那么使用 K-means 会有什么样的问题呢？

K 值敏感

首先是社团数目的问题。在进行 K-means 聚类过程之前，我们曾假设整个社会会形成 K 个社团。然而这样的假设却很经不住推敲，社团数目是由整个社会中所有成员的特质与个性共同影响决定的，因此 K-means 中的 K 本身就是这个算法的一个弊端，聚类后类别的数目需要提前指定。

初始值敏感

其次，为这 K 个社团指定 K 个中心人物的做法也有值得诟病的地方。在 K-means 中，我们采取的是随机选取中心人物的做法。随机会使得选取十分不固定，有可能满足，也有可能并不满足社团对中心人物的要求。对于中心人物并不具备代表性的情况，后加入的成员就仿佛是跟错了人，站错了队。对于中心人物间相互区分度较弱的情况，后加入的成员可能会感到迷惑，最终导致本该属于同一个社团的两个成员，由于社团间并不显著的差别而被硬生生分到两个类别中。

离群点敏感

最后出现的情况则是我们在 K-means 中最不愿意看见的，这个情况可以使用

一句谚语来概括——一粒老鼠屎坏了一锅粥。在我们进行社团聚类的过程中，由于整个社会中的每一个个体都要被强制安排到某个社团中，这极有可能造成原本并不和任何社团相似的成员也需要做出选择。学术上称这样的成员为离群点。由于我们在理解社团的属性和特性的时候，是以该社团中所有成员的特性来进行平均的，因而这样的离群成员越多，这个社团在外界眼光中的定位就越偏离其本来的初衷。

总结来说，盲目指定社团的数目、随机指定初始中心人物以及若干存在的异类都会造成 K-means 聚类算法的不科学。

9.7 层次聚类：分而治之与抱团取暖

说到层次聚类，还是要先来看看最能够反映这个算法特点的两个字“层次”。我们通常形容一个人具有较高的视野与宽广的胸怀时，会说这个人“层次高”。而想表示一个人位高权重的时候，也会使用“层级高”来形容。层次或者层级所表示的就是某种等级秩序。例如思想家和教育家孔子就曾经把人划分为“庸、士、君、贤、圣”五个层次，每一个层次都比上一个层次拥有更高的等级。由于等级更高，因而能够达标的人也更少，即层次越高的等级成员数目越少。

与上述例子类似的还有地理上的归属地概念。各种地理维度形成了一个层次分明的框架，仿佛是一个金字塔，这个金字塔共分为六层，最顶层是大洲，依次向下是国家、省、市、区县、街道。越处于金字塔下层的地理维度越具有更多的成员，每个下层的成员都隶属于它上一层的某个成员，例如合肥市隶属于安徽省，而安徽省又隶属于中国。

层次聚类便是这样一种让成员在每一层形成不同聚集的方法。仍以上述地理层次框架和地球上所有的成员为待聚类的对象。在大洲这个顶层，所有成员被聚集成七类；而到了国家这个层面，所有的成员则被聚集成 200 多个类。正如层次等级秩序所反映的，层级越向下，类别越多；层级越向上，类别越少。即便是这样的层次聚类，也有两种不同的方式，这两种方式一种从上向下进行，另一种则反向进行，我们分别冠以“分而治之”与“抱团取暖”的名称加以诠释。

9.7.1 分而治之

分而治之本是一种军事谋略，讲的是将体量巨大的敌人化整为零，从而各个击破的方法。在《孙子兵法》中有“十则围之，五则攻之，倍则分之，敌则能战之，不若则能避之”，讲的便是这样的方法。

计算机科学总是擅于向其他学科借鉴，特别是社会学中的理论。依据此，计算机科学家们也发明了一种分而治之的计算机算法。其思想和军事思想差不多，主要是将大问题划分成许多可以被解决的小问题，通过解决小问题从而使得大问题得以解决。那么在层次聚类中，它又是如何具体体现的呢？

在一个由 16 个人参加的会议上，我们需要把与会者划分为 4 类。使用分而治之的策略可以如下操作。首先在最高的层次上，我们认为这 16 位与会人士是一个不可分割的整体，他们属于同一个大的类别。接着，我们可以通过寻找到某种特征将这 16 人分为两类。这里的特征可以是衣着的颜色，或者是否戴眼镜，或是性别。选取什么样的特征取决于具体的情况。经过一次判别，原来的一个大类分裂成了两个较小的类。每个类别中的成员数目可以相等也可以不等。为了分成 4 个类，我们可以继续针对每个类再次选择某个特征进行划分，从而得到最终的 4 个类别。

上述获得最终聚类类别的形式是从上向下进行的，即由一个大类逐步获得各个小类。最初的类别不断分裂，直到得到的聚类类别数目已经满足我们的要求，或者不能再分裂为止。

9.7.2 抱团取暖

既然分而治之的聚类是自顶向下的，那么自然地会联想到抱团取暖的聚类是自下而上的。

以曾经公司团建时的一个游戏为例来解释抱团取暖。游戏的名称是“一元五角”，规则是这样的：在团队中女生代表一元，男生代表五角，游戏开始时大家随机游走，当主持人报出一个金额之后，大家需要快速找到合适的同性或异性队友组合成这个金额，并通过拥抱的方式确定组合的关系，以防止其他成

员插入。组合完成后，落单的成员若无法再次组合成主持人规定金额就被淘汰出局。

在这个游戏最开始的时候，我们可以把每个人看成一个类别。在进行了一次抱团组队之后，若把拥抱在一起的人看成一个整体的话，整个参与游戏的类别就减少了。这种原来聚类类别数目繁多且细碎，经过结合而减少的方法就是一种抱团取暖式的聚类。

这种方式从下开始向上进行，即通过每一个细小的成员相互抱团而形成小组或是更大的组织，从而达到聚类的效果。抱团取暖的规则我们犹未可知，但在真正的聚类中我们却可以通过度量各个类别之间的相似程度来聚合两个或者更多的类。

类比的描述并不一定很精确，但只要你开始了这样的思考方式，就根本停不下来。这种一点点简单联想，触达和联通若干门科本质的方法，使得我们对于本行的技术可以驾轻就熟，也使得我们在面对洪水猛兽般的未知时不再踟蹰。

9.8 DBScan：帝国崛起的定居、建国与扩张

对于聚类，还有什么妙招呢？本节要聊的是一个名为 DBScan 的方法。看到 DB 首先想到数据库，这是 IT 人员的本能，可是这里的 DB 并非是 Database 的意思，而是 Density-Based 的首字母缩写，即基于密度的含义。Scan 字面意思是扫描，而实际上 Scan 是 Spatial Clustering of Application with Noise 的首字母缩，即噪声应用空间聚类。连在一起看，这个方法大致就是一种密度聚类方法，并且能够很好处理噪声问题。

密度通常指的是某一空间单位中的量。这样的空间可以是长度、面积或者体积，而其对应的量则可以是数量、质量、电量等。例如我们在物理学中说的密度更多的是单位体积下的质量，在城市规划设计过程中说到的密度则指的是单位土地面积上的人口数量。

使用密度来进行聚类相比之前的诸多聚类算法有着无可比拟的优势，而且使用密度进行聚类的思想与社群的建立与发展更加相似。

9.8.1 密度打败划分

前面介绍过的 K-means 也被称为划分聚类，这个名字很形象，即通过确定若干个初始的中心人物，从而将剩下的新成员划分到各自的类别中去。层次聚类则是通过自顶向下或者自下而上的方式来对待聚类的成员的。那么密度呢？

先来看看下面的例子。图 9-5 的三幅图中，最左侧的图片中散落了一些点，这些点都是待聚类的成员。仅通过观察，我们很容易识别出图中相互倒扣的两个回行飞镖的图案，因此聚类的结果应该更倾向于将属于同一个回行飞镖的点归到一类。



图 9-5 聚类示例图

然而事实是怎么样呢？我们使用之前介绍过的 K-means 算法和这里即将介绍的 DBScan 算法分别对左侧图中的点进行聚类处理，得到了右侧的两幅图，中间和右边的两幅图片中各有两类。

中间的图正是 K-means 聚类得到的结果，而右边的图则是 DBScan 聚类得到的结果。从图中不难发现，K-means 由于在聚类过程中把点安放到与自己更相近的类别中（在数学上表示为距离上更近），于是极有可能出现飞镖相互紧邻的边缘难以判定。相比较而言，DBScan 在解决这类问题上更有优势。究其原因，是因为 K-means 聚类过分依赖于点与点之间的距离度量，忽略了周边的因素，而密度聚类则考虑了周边的情况。

由此可见，密度聚类对于一些容易纠缠不清的临近点有很好的处理效果。除此之外，密度聚类方法也可以解决离群点的问题。在密度聚类中，由于离群点远离大部分的成员，故而在其广阔的周边并不存在其他成员。按照密度聚类的思想，其周围空间中点的数量很少，所以密度自然也很小。这个时候我们没必要像 K-means 一样，一定要为其指定一个已然存在的类，而是可以通过标记其为离群点而忽视它，这样也使得已经聚成的类别不会受到离群点的干扰。

9.8.2 相似的帝国发展路径

为了简化 DBScan 算法的过程，我们将其和政治中的帝国建国步骤相类比，得出了一个能够大概描述 DBScan 算法步骤的三部曲：定居、建国、扩张。

亚历山大帝国又名马其顿王国，这是由其地理位置而决定的。早年的王国地处希腊东北部的马其顿地区，在波斯和希腊之间的“波西战争”时依附于波斯帝国。由于波斯和希腊双方消耗甚多，给了马其顿帝国以崛起的机会。亚历山大的父亲腓力二世主持当时的国家，趁希腊内乱之际入主希腊，成为了希腊各城邦推崇的霸主。到这个时候算是建国完成。亚历山大接手后，马其顿王国开始扩张征程。先后征服了黑海周边、波斯湾、非洲的埃及，乃至东方的印度。从马其顿定居点走出，到希腊建国，最后扩张成为一个横跨亚、非、欧三个大陆的帝国，不可谓不伟大。

在世界历史的版图上，一个国家可以看成一个聚类后的类别。首先选择一个（定居）点，如果其足够强大（围绕其有足够多的点），那么就建立国家（形成类别），将已经纳入这个国家的成员当成国家（类别）的一份子。如果该国成员周边也有足够的点能够满足建国的条件，就将那些点也相继纳入这个国家中来（扩张）。这样就可以获得整张数据点版图上的国家（类别）数目以及大小了。

若能用上面的概念模式来类比学习 DBScan，便再好不过了。在 DBScan 中有两个很重要的概念，一个叫作周边距离，一个叫作周边距离内的密度。周边距离可以理解为以某点为圆心画圆的半径长度。成语“鱼肉乡里”中的“乡里”指的便是政权所掌控的地方。此外，周边距离中的密度则指的是这个空间范围内的人口密度。要满足建立一个国家或者王朝的条件，首先需要满足的是人口条件，即在一定的周边距离范围内，人口密度需要达到一定的数值。DBScan 中也有这样的要求，如果某个成员周边距离范围内的成员达到了一定的数目，则他们应该属于同一个类别，这里的类别便对应于古时的一个王朝或国家。

9.9 关联规则挖掘：“啤酒和尿布”是个谎言

让我们继续沿着数据挖掘的道路向前进发，本节要介绍的是关联规则。很显

然,从名字上看,关联规则似乎无法让人直接理解,那么关联规则究竟有哪些有趣的故事,我们又该如何使用这个算法工具呢?

9.9.1 讹传已久的商业故事

为了解释清楚什么是关联规则,我们需要从一个业内讹传已久的数据挖掘商业故事说起。这则商业故事的重要性可谓举足轻重,甚至可以说,尽管关联规则并不是数据挖掘的全部,但是对行外人来说,这个案例却是数据挖掘对他们而言的全部。

关联规则之所以这么容易为人们接受,是有原因的。首先,它嫁接的故事比较巧妙。在当时,超级市场开始风靡,借助热点话题和新兴事物容易激发人们的兴趣。其次,它的内在本质是探究因果关系。人类从存在以来就一直渴望探究因果关系。例如狩猎时期人们找寻到猎物迁徙的规律,农耕社会找寻对植物生长有利的因素,工业社会探究社会发展的铁律,信息社会挖掘事物背后的原因等,其本质都是在追求一种确定性。关联规则就提供给了人们找到因果性的途径,因而也受到大家的喜欢。

说了这么多,想必你已经猜测到了,这个故事就是风靡世界的“尿布与啤酒”。这个故事发生在沃尔玛(又一说法是 Target 公司)超级市场中国。在超市的卖场收银结算数据中,工程师们发现,经常买尿布的顾客也会顺手捎带一罐啤酒。这一现象引起了人们的好奇,通过探究其背后的原因,销售人员发现在国外买尿布的大多是父亲,他们也喜欢喝酒,于是会同时采购两者。卖场挖掘出这样一则规律后,就可以通过摆放货架而促进交叉销售,从而提升业绩。至此,一个数据驱动商业决策的故事就陈述完了,十分完美。从数据的采集(消费者购物清单)到算法挖掘(关联规则),再到原因探究,最后到指导商业行为(货架摆放,交叉销售),一应俱全。

但是事实情况如何呢?这则故事背后有许多解读版本,大致有以下几种。

数据 or 经验

这个故事已经流传了许多年,现如今很多仓储型超市也并没有一套完整的数据挖掘促进交叉销售的解决方案。这并不是因为数据挖掘无法挖掘出交叉销售的

商品，事实上，利用关联规则确实能够找到一些结论。但是通过数据挖掘得到的交叉销售带来的利润提升还不一定能够覆盖为了实施数据挖掘项目所付出的成本，自然商家要思量再三。我们现在看到的货架陈列变动以及小商品收银台处的顺手捎带货架，大多数还是依靠销售人员的经验或是消费心理学的研究成果来确定的，其中或许并没有太多计算机算法或大数据技术在帮忙。

来源的谬误

这个故事中以沃尔玛超市或者 Target 超市为载体，极大可能是为了借助这两个全球知名超级市场的品牌而使这个故事的传播更具可信性。但这个案例的来源并非沃尔玛，而是一家药品商店 Osco Drugs，数据挖掘的工作确实有所开展，也确实是基于真实数据，但是被过分夸大用于商业宣传，其目的只不过是某工程师为了销售自家公司的行业软件罢了。

因果与相关

人们倾向于接受的因果性也许本身就没有什么意义。我们得到的因果性往往具有这样的结构：如果 A 那么 B，这等于给出了一个预测未来的模式，只要出现了 A，我们就可以断言 B。然而百密一疏，导致 B 的原因也有很多，盲目相信 A 这一种原因有可能会使得我们一叶而障目。即便因果性是由专业人士找出的，也不够客观，更别说别有用心的商家与技术人员为了操纵大众意志而给出的伪因果结论了。

商业的讹传也罢，真实的场景也好，故事终归是故事。正如因果性中一个果对应多个因一样，在评价关联规则的时候我们也应该秉持多元的价值观。一些项目未能真实地实施，可能是商业上基于投入产出比的考量，而并非算法有问题。

9.9.2 关联规则的三重门

既然算法没有原罪，那我们就有学习和了解的必要。关联规则的一个核心目标就是找出“如果 A 那么 B”的规则形式。在发现规则形式的过程中，算法为我们设置了三重障碍，只要突破这三重障碍，就可以达到最终的目标。

共同出现是基础

因果或者相关的共同基础是事物共同出现。可以从两个形影不离的异性判

断出亲密关系，也会把霉运和一些潜在原因联系到一起。这都是因果性和相关性最好的例子，归根到底是在一个特定范围内同时或者近乎同时出现的事物间寻找原因。

因此第一重门就要求我们学会在整体中找到两个事物或者多个事物之间共同出现的次数与频率。以超市购物为例，每次消费都是一次交易，每次交易中顾客可能会购买不止一件商品。如果有 1000 张购物小票，其中出现食用盐的有 200 张，出现大米的有 100 张，而食用盐与大米同时出现的有 50 张，我们依据这些数据就可以计算出食用盐和大米共同出现的频率是 5% ($50/1000$)。学术上称这个数值为支持度。

谁因谁果想清楚

找到了共现就有了找出因果或者关联的基础。可是究竟是买食用盐的人更倾向于采购大米，还是买了大米的人会顺手捎带一包盐？这个问题决定了大米和食用盐谁是因谁是果。在进行这两个问题判断的时候，无非就是在“若买大米则买食用盐”和“若买食用盐则买大米”之间做选择。这里共现的事物只有两个，因此我们遇到可供挑选的情况也只有两个。共现的事物越多，那么需要判断孰因孰果的情况就越多。这和我们在复杂事物中抽丝剥茧分析原因所遇到的复杂场景是类似的。

在上述例子中，我们已经获得了“买大米且买食用盐”的共现频率，那么怎么样才能对“若买大米则买食用盐”的情况进行计算呢？很简单，“若买大米则买食用盐”可以翻译为在所有购买大米的交易中购买食用盐的交易所占的比例。我们需要计算出包含大米的交易出现的频率，此项为 10% ($100/1000$)。而在这个情况下还购买食用盐的交易则就是我们之前已经计算出的既买大米也买食用盐的交易出现的频率，即 5%。使用上述数字相比，就可以得到 50% 的结果。学术上称这个值为置信度。同理，我们也可以计算出“若买食用盐则买大米”的置信度为 25%。因此我们更倾向于选择买大米作为原因，而将购买食用盐作为结果。

找到规则要靠谱

找到规则只能算是第二重门，考究这个规则靠不靠谱则是第三重门。以刚刚找到的置信度为 50% 的规则“若买大米则买食用盐”为例，究竟这个规则是否靠谱，需要设立一个衡量靠谱与否的标准，这个标准需要以结果为考量。

购买食用盐发生的频率是 20%，而在购买了大米之后发生购买食用盐的频率提升为 50%，由此可见这个规则还是有提升效果的。

在学术上，我们使用两者的比值（ $50\%/20\%=2.5$ ）是否大于 1 来判断其是否靠谱，这被称为提升度。情况也并非总是如此，如果在 1000 宗交易中出现食用盐的交易数量变动为 500 以上，前面的支持度与置信度并没有变化，但提升度则有所改变，于是刚才的规则就似乎变得不再靠谱了。

说到这，不知道你有没有一点感触，我们说的因果性不过是对一个规则的相信程度比较高而已。也就是说，并非任何因果性都是 100%靠谱的，所以多了解一些科学背后的原理对于我们去伪存真，培养批判思维精神大有裨益。

9.10 时间序列分析：聊聊《周易》

时间序列是我们即将介绍的最后一个具体的数据挖掘技术。

时间序列分析一看名字就知道是做什么的。使用这个方法一定会涉及时间因素，当时间全部排列开形成一个序列的时候，就会有规律可循。这个规律在我脑海中首先映射出的是 \sin 和 \cos 函数的波浪式图像。时间序列分析算法显然不是简单用三角函数就可以概括与言明的，下面从玄学的角度聊聊时间序列，然后再转回严谨的科学算法。

9.10.1 时间序列分析的玄妙

先以一个例子来认识时间序列分析吧。

图 9-6 中总共包含四幅子图，不妨从上向下标记为 1~4。第 1 幅图是实际的数据情况，我们称之为现实。图 2~4 分别是对图 1 的分解，意味着图 1 可以拆解成下面 3 幅图。图 2 描述的是图 1 的简单趋势，是一种大方向；图 3 体现的是图 1 的规律性；图 4 则反映了噪音数据，变化莫测，难以捉摸。时间序列分析就是从现实中找到趋势、发现规律、剥离噪音。

历史大势，浩浩汤汤。谁能把握趋势，谁便能成就一代伟业。因此在中国古代，历朝历代的君主也好，平民百姓也罢，都希望能够顺势而为，既为代代相传，

也为当世福乐。从集权统治的中心到市井民间，玄学找到了生根发芽的土壤。周文王姬昌据说是《周易》的作者，该书中提到了三种“易”，一曰简易（大方向），二曰不易（规律性），三曰变易（变幻莫测）。这三种易对于现实社会的刻画正是现如今时间序列分析的雏形。要知道《周易》的发明比时间序列分析算法的发现足足早了近 2000 年。

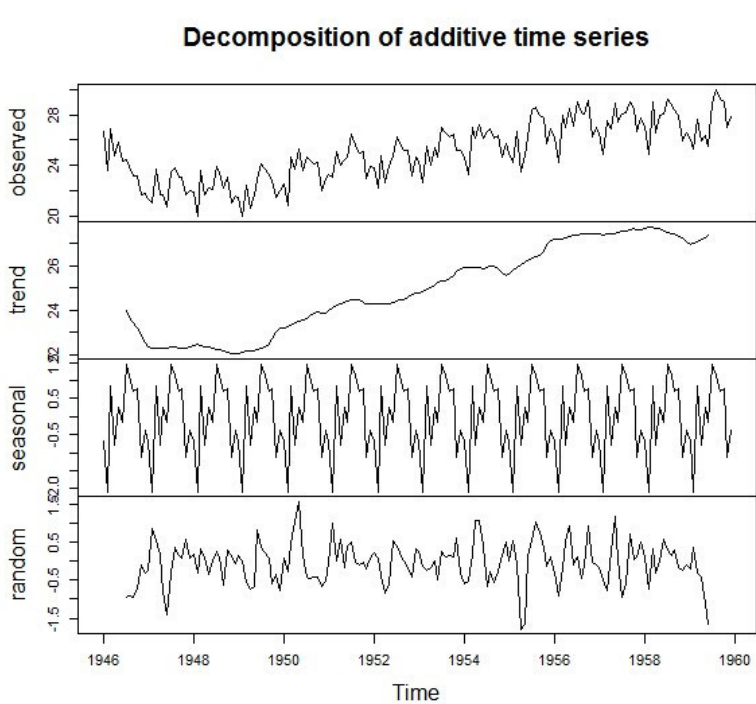


图 9-6 时间序列数据的分解示意图

引申思考，我们研究数据的价值，实际上是在治理数据，而《周易》则是在探究一种治人治社会乃至治国的方法谋略。《周易》的“周”“易”二字，可以诠释为“周人、周全、周转”“简易、不易、变易”。这六个词所反映的是周人产生《周易》的事实，这项理论既是对社会周全的描述，也是对社会周转运行状态的总结。而社会中所蕴含的道理则是简单的，其中既有恒定不变的定理，也有情随境迁的动态变化。

数据治理、探究和挖掘也可与之对应，分别是“产生、全面、动态”“真实、规律、噪声”。数据从业务场景中产生，具备了全面多样化（Variety）和动态变化

(Velocity) 的特点。数据是真实世界的反映，我们需要从噪声中找出趋势与规律。在这种层面上，两者又被高度统一起来了。

9.10.2 时间序列分析的正经

时间序列分析算法有很多种，任何一种中都涉及诸多新概念，其对数学的要求也并不低。而我的出发点则是在只有高中数学知识的前提下理解数据挖掘技术，因此我会选取一个算法，在模糊易理解和精确更科学之间找到一个平衡点来介绍。

Box-Jenkins ARIMA 是很著名的时间序列分析算法。Box 与 Jenkins 是 ARIMA 的发明者，这样我们就可以简化这个算法为 ARIMA。这个名称还是很复杂，明显是一些英文单词首字母的缩写，那么这些词究竟代表什么呢？我们不妨先做拆分，将 ARIMA 拆成三段：AR、I 与 MA，分别介绍。

习以为常的 AR

AR 是 Auto Regression 的缩写，即自动回归。它和之前介绍的线性回归是有一定关联的，两者都是用于预测的。这里的 AR 可以单纯地认为是前面的线性拟合，不再赘述。

差强人意的 I

对于 I 来说，它是 Integrated 的缩写，本意是链接 AR 与 MA，也就是将两者联系起来的一种方法。而我在这里更想赋予它另一层含义，即 difference。这个单词的英文释义是差别。在这个算法中，学术名称是差分。

还记得我们以前学过的等差数列吗？首项 a_1 ，末项 a_n ，公差 d ，通项公式 $a_n = a_1 + (n-1)d$ ， d 就是两项的差值。学术上称这样前后相邻两项的差值为“1 阶差分”，仿佛是把前后相邻的数据放在了一个台阶上，然后求出相邻台阶上数据的差值。那么“2 阶差分”是否就是求出跨两个台阶的数据差值呢？并非如此。2 阶差分是在 1 阶差分的基础上进行的，是对 1 阶差分的结果再求 1 阶差分，所以其中的“2”表示的是求差分的次数。

举个例子吧。数列 (2, 4, 6, 8, 10) 里面包含两个趋势，一个是向上的（递增），一个是平稳的（每次递增相同）。当使用 1 阶差分（后项减前项）计算后即得到数列 (2, 2, 2, 2, 2)，这个时候数列中的数据已经趋于稳定，便于研究数

列性质。这里列举的等差数列只不过是简单和粗浅的例子，实际收集的数据上下波动，难以像该数列一样具有一眼看穿的性质和趋势，因此使用差分就可以使波动的趋势减缓，数据的变化趋于稳定。你可以随意写下一组数据，然后使用差分试一试，看看是否能够达到稳定数据的效果。没有什么是差分不能解决的，如果不行，那就再做一次差分。

对于拟合（AR）来说，它善于处理规律性和趋势性的数据，而为了把按时间排列的序列数据流转换成规律性的数据就需要差分与 MA 过程的帮助。经过差分之后，数据变得更加稳定，便于研究趋势，可是基于处理后的数据得出研究结果之后又如何还原成原来的数据呢？其实只要知道原始数据以及做了几次差分，就可以恢复原数据。

圆滑当道的 MA

MA 是 Moving Average 的缩写，即滑动平均。其过程是使用滑动窗口顺次地处理窗口内的数据，最简单的方式就是求这个窗口内数据的平均值。例如数据（2，4，6，8，10）取滑动窗口大小为 3 并从最左侧开始处理，则变成（2，4，4，6，8）。该数列相较原来数据序列更加趋于稳定。其主要作用和差分是类似的，即将数据变得更加温柔与平和。

我们平时使用的美图秀秀中的磨皮功能就具有滑动平均技术的影子，其基本原理就是将面部色素沉着所在位置的颜色与周围一定范围内的皮肤颜色取一个平均值，替代原色素位置。真实算法更为复杂，但大体思想如此。

了解了 ARIMA 的三个过程之后，我们便可以将其综合起来了解具体的时间序列分析是如何进行的。实际过程是先 MA（滑动平均），再 I（差分），最后 AR（回归预测）。总体来看，在使用时间序列分析算法时经历了先稳定数据，再找到规律的两大过程。

9.11 集成学习：三个臭皮匠赛过诸葛亮

从本节开始，我们介绍的内容就不能称为算法了，而是一种手段或方法，其目的是为了研究一类问题。在我个人看来，算法与方法的本质区别大致有两点：第一，算法是一套清晰流程，有章可循，而方法则是一种笼统的思路，只给方向

性的指导，其内部可能包含多种算法的实现形式；第二，算法相对来说层次较高，大概只有 IT 领域的人士才会用到这个词汇，而相对来说方法就比较平民化，对于大多数人来说，每个人都可以针对当下的问题找到一种属于自己的解决方法。

本节要介绍的集成学习，以及接下来的若干节，都将带领我们这艘数据挖掘的探险之船驶入比算法更深的海域。准备好，我们要再度启航了。

9.11.1 多拜师与拜大师

在奥运会中最早拉开帷幕的项目总是最受人瞩目，中国队历年的奥运金牌开门红也总是由国家射击队带来的，我们就以射击这件事来看看什么才是一个好的射手。

图 9-7 是四位射手射击训练的靶纸，作为教练的你，现在要挑选两名选手参赛，你会怎么选择呢？毋庸置疑，左上方的靶纸全都正中红心，这名选手首先被选中。右下角的选手是最差劲的一个，也自然最先被淘汰。悬念留在了打出右上角和左下角两张靶纸的选手身上。

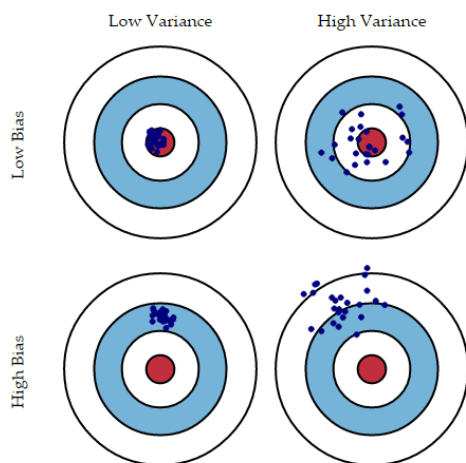


图 9-7 方差 (variance) 与偏移 (bias) 示意图

图片来源：

<https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning/>

相较而言，右上角的选手基本上都打在中心附近，但是分散范围较大，因此成绩不稳定。左下角的选手每次打靶都能打出大致相同的成绩，但是总也打不到靶心附近。如果把中心看成实际目标，而将打靶看成不断接近和预测实际目标的过程，那么数据挖掘和机器学习算法做的就是这件事。我们总是期望算法表现得像最早入选的左上角选手一样，但很有可能算法和最早淘汰的右下角选手一样，水平很差。

我们常听到数据挖掘工程师和算法工程师提到“训练模型”，正如射击选手需要训练一样，算法也需要训练，从而使得打靶成绩从右下角变成左上角。但是事与愿违，很多情况下算法和射击选手会落入另外两种无奈的情况中。

为什么会出现这样的情况呢？在学习和体育训练中我们往往会遇到一个情况，从菜鸟到一般水平相对容易，就好比从 60 分提高到 90 分较容易，而从一般水平再向上精进就十分困难了。90 分的情况对应的就是打靶中的左下角和右上角的情况。一个运动员始终跟随一个教练学习，即便是青出于而蓝胜于蓝，到达一定的阶段后，也很难再提升。这就是为什么一个知名运动员往往需要由国内外多个教练指导的原因，多拜师的好处也就因此而显现。

多个教练能够带来多种不同的训练风格和训练方法，有助于互相补充，帮助运动员弥补短板，从而实现二次飞跃。同样的道理运用在数据挖掘和机器学习算法上也是一样的，当我们综合不同的算法针对性地在数据上解决一个问题，或者使用同样的算法从不同侧面解决数据中的一个问题时，都可以对问题解决的效果有很好的提升。这种情况下，无论是多种算法，还是多样数据，都是利用三个臭皮匠的智慧去解决本由一个诸葛亮才可解决的难题，学术上称为“集成学习”，是集合了所有的智慧来进行训练以提升学习效果的方法。

9.11.2 向大家与失败学习

先从字面意思来看“集成学习”。所谓集成就是把一切能够找到的东西收集起来，然后捏合成某物。譬如软件集成就是将多个软件模块放在一起针对个性化需求进行定制化开发或高级功能研发；系统集成就是将多个系统（如监控系统、软件系统、硬件系统等）集合到一起进行开发并投入使用。

而我们在这里要说的集成学习又是怎么回事呢？很简单，集成学习就是把多个学习结果集成起来。这里的学习结果可以从空间和时间上分成两类。

向大家学习

针对一个问题，不同的人会有不同的解决方法，相似地，不同的计算机算法也会有不同的解法，我们称某个计算机算法解决某个问题的结果为学习结果，仿佛是计算机具备了人类的智慧一般。可以将同样的数据扔给多种计算机算法，看看会有什么样的学习结果；或是从一份数据中采样获得多种不同样式的数据，然后让相似的计算机算法进行学习，也会得到不同的学习结果。将这些学习结果进行汇总，并通过加权或者投票的方式得到最终的判断标准。利用这个方法，我们将平行空间上的学习结果进行了集成，学术上称为 Bagging。

Bagging 是 Bootstrap Aggregating 的缩写，Aggregating 有集合在一起的意思，简单地理解，bootstrap 由两个单词构成，boot 是靴子的意思，另外 strap 有捆绑用的带子的意思，既可以表示鞋带，也可以表示捆绑用的装束。因而我们可以推测，鞋带的作用是帮助靴子紧凑不至于因为松松垮垮而脱落。另一方面，它能够起到的作用是较小的，本质上还是要靠鞋子自身的形状附着在脚上，因此我们说 bootstrap 有自助的意思。

Bagging 方法的发明人就是在前文中提到的发明了决策树 CART 算法的里奥·布莱曼（Leo Breiman）。如果在数据上构建决策树，并使用 Bagging 方法，就得到了布莱曼一生中的第三个发明成就“随机森林算法”。这个算法预示着，在随机数据上通过多棵树构成的森林算法来进行学习，十分形象。

向失败学习

向大家学习的方法是借助同一时间不同空间中多人的合力来获得最终结果的，而这一结果也可以通过向自身犯过的错误进行学习来获得。我们通常说“失败是最好的老师”“失败是成功之母”，讲的都是从失败中学习的重要性。由于失败，所以我们会更加关注错误和失误的案例，从而在下一次遇到的时候提高警惕以避免再犯。

计算机也具备这样的“反思”能力，只不过它是通过在每次学习的结果中找出错误的案例，给予错误案例更高的权重以进行下一次学习的。最后通过将不同

时间点的自我学习结果集合起来，通过加权或者投票的方式，以获得最终的结果，这便是计算机向自我失败学习的方式。学术上称这样的方式为 Boosting，有增强和推进的意思。

从策略上来说，集成学习是一种虚心纳谏、智慧决策的方法。无论是 Bagging 还是 Boosting，一个是提升，一个是推进，两个集成学习方法都有使整体变好的趋势，这也恰恰反映了集成学习的根本目的。

9.12 文本挖掘：让机器读懂你

如果说集成学习是一种把各种学习结果集成到一起的方法，那么本节要介绍的文本挖掘就要针对文本这类具体的数据对象使用算法玩出一些花样。在行外人看来，数据不过是阿拉伯数字。诚然，大多数情况下，数据的呈现形态是阿拉伯数字，但是像文本、图像、视频这一类对象在计算机看来也是数据，因为这些对象在计算机硬盘上的存储形式与简单的阿拉伯数字存储并无差别。所以，文本也是一种数据。

顾名思义，文本挖掘则是要挖掘文本。文本是大数据领域数据源的一种。简单理解起来，可以把文本理解成文字，也就是微博发的各种状态，以及在各种论坛和博客下面给别人的评论等。这里提到的可被挖掘的文本除了自己发布的内容，还包括转发、收藏的新闻与文章等。用一句不太严谨的话来说，文本挖掘就是通过计算机方法找出一切文字中蕴含的信息，无论这些信息发布在什么地方，由谁发布，在后台是如何存储的。

对于文本挖掘来说，行内有一些典型的应用场景和技术。了解这些技术的应用场景可以扩展数据产品经理日常工作的思路，也为丰富数据产品形态奠定了基础。这些技术从分词到主题发现，从实体识别到文本分类，从摘要到情感分析，一应俱全。接下来就分别介绍。

分词

分词是汉语文本挖掘的特有技术，为什么这么说呢？以英文句子 “I have an apple” 为例，这里面的每个词都可以使用空格隔开，然后通过对每个词意思的识别来了解这句话的意思。而对于中文语句 “我有一个苹果” 来说，尽管很容易把

这句汉语拆开为“我、有、一个、苹果”，但是却很难对每一句话找到一个通用的规则来进行切分，更别说使用计算机来识别了。特别是当一些特殊句子出现的时候，譬如“安徽省长江流域”，其中“长”既可以和后面的“江流域”合起来构成一个词汇，也可以和前面的“省”合在一起构成一个词汇。计算机遇到这样的情况又如何处理呢？在中文里，计算机用来处理上述问题的技术就称为分词。

分词技术是所有文本挖掘技术的基础。目前已经有一些工具可以做到大致准确地划分大众型的文本，如果需要处理一些特定领域的文本，人工构建一些专用的行业词库导入原有的通用型分词模型便可以提升效果了。分词之后的一个最简单的应用就是进行热词统计，其本质就是数数，看看什么样的词出现的频率较高，出现频率较高的就是热度高的词汇。同时，我们可以利用一些可视化软件绘制词云图来展示热度词汇。

主题发现

如果你听到身边人的谈话中蹦出“赵薇”“高圆圆”“刘若英”“演唱会”“电视剧”等若干词汇，其他的并没有听清楚，你会认为他们在谈论什么话题？如果你的答案是娱乐或是其他类似的主题，那么恭喜你，你已经学会如何总结概念了。

话题就是谈话的主题，主题可以是谈话代表词汇中的一个或几个，也可以是把这些词放在一起后经过人工提炼而得出的某一个抽象的概念。你或许认为，将文本分词后，直接将关键词拿出来就可以作为一段文本的主题了。没错！可是什么是关键词？难道是热度词汇吗？如果是这样，那么大部分情况下你得出的主题应该是“的”“地”“得”“你”“我”“他”，然而这些并不是什么主题，也并不能够带给你任何的信息量。

因此，主题提取需要一个专门的方法，这种方法叫作隐狄利克雷分布（LDA）。对于数据产品经理来说，你仅需认识这个方法，而并不需要知道其内部机理。该技术相对较为成熟，可以应用在用户兴趣标签提取等场景中。

实体识别

从学术的角度来说，实体识别称为命名实体识别（Named Entity Recognition）会更加合适。要想明白什么是实体，一个例子便可以言明。“我爱北京天安门”这句话中有三个实体，分别是“我”“北京”“天安门”。实际上实体就是真实存在的

个体，它可以是人物、地点、建筑名称等，只要真实存在即可。除此之外的专有名词也可以算是实体。

实体识别基于分词技术，主要应用在机器翻译和问答系统中。在机器翻译中，我们需要把天安门广场翻译成 Tian'anmen Square，这就需要首先识别出天安门广场这个实体。在问答系统中，特别是人工智能时代的聊天机器人应用中，也有实体识别的身影。当你和聊天机器人谈论美国大选的时候，它需要识别出你话中的对象，然后才能从它的知识库中加以搜索并作答。当你咨询它某地的天气时，它也需要首先识别出地点实体，才能够给出正确的提示。总体来说，实体识别主要是识别实物、时间与数字类的内容，并在具体的场景中应用。

文本分类/聚类

分类和聚类已经在前面介绍过多次，当文本成为对象时，分类和聚类中的点与成员就变成了文本。因此，文本分类和聚类就是要把相似的文本找出来放在一起。

文本分类最典型的代表应用是对新闻的分类，譬如在今日头条等内容分发类软件和网站上通过分类算法给各种新闻内容打上标签，从而自动归档到相应的新闻类别中。

下面列举两个文本聚类的主要应用。一是文本摘要，通过对相似的文章进行聚类，从而消除冗余、融合信息并最终生成能够代表这一类新闻或文本的摘要信息。二是搜索引擎中的去重，用户在搜索一个关键字的时候，难免会有多篇相似的文章。对于用户来说，搜索引擎应该在第一屏中提供更为多样化的结果，而不是任由相似的文章占据用户的视野。通过聚类就可以把相似的文章聚集起来，用户就不会查看到冗余和重复信息了。

情感分析

一篇文章或者一句话，究竟是褒还是贬？褒贬的程度是深还是浅？判断这些问题的过程就称为情感分析。目前的情感分析是舆情分析应用中必不可少的技术，通过搜集网友关于某人某事件的评论，就可以分析得出大众对于该人该事件的正负面评价，用以指导决策。

除了官方针对舆情的分析场景外，商家往往也会关注用户在团购点评类网站

的反馈信息,其处理过程既使用了用户的投票信息(如外卖 App 中的星级打分等),也使用了用户点评文本中一些情感词的倾向性分析。

以上的技术与场景是文本挖掘中最为常见的应用。技术一直在发展,各行各业涌现出了许多叠加在上述场景和技术上的新应用。在日常工作中积累这些新的应用,了解哪些技术是可能的,哪些技术是不可能的,可能的技术中有哪些是成熟可落地的,哪些是仅限于前瞻 demo 演示的,对于数据产品经理来说,这样便可游刃有余地规划靠谱的数据产品了。

9.13 社交网络：隐私无处遁形

Web 时代从 1.0 迈向 2.0,进而挺近 3.0,最大的不同便是人获取信息的途径。在信息的元生代,信息就在那里,人主动地去寻找信息,例如我们从一些门户网站看到新闻,也会偶尔给予些许点评。随着网络形式的多样化,论坛出现了,人们可以在一个社区内自由发言。随着人与人沟通刚需的演进,社区论坛进化为社交网络,原来单一的获取信息与给予点评变成了互相递送信息以及无所不在的沟通。

如果说文本挖掘更多的是对前一个时代信息的处理,那么本节要讲的社交网络挖掘则是在其基础上与时俱进的演化。对于社交网络,没有人陌生,从开心网到人人网,从微博到微信,我想大家至少应该体验过其中的两三样产品吧。社交网络沉淀的数据究竟能够带给我们什么价值,我们能够基于它们做什么呢?和前文一样,这里列举几个经典的应用,供数据产品经理参考。同样地,数据产品经理只需要知道这些应用的输出以及成熟度即可,剩余的就交给团队中的工程师和算法同学吧。

影响力发现

影响力发现是指在社交网络中发现有影响力的人。在现实生活中,有影响力的人一般具有如下特性:身居要位、言论有影响力、圈子高端。在真实的物理世界中我们似乎很容易识别这些人,那些企业的法人、政府的高官、社会的名流就是这样的人,可是计算机究竟要怎么完成这样的任务呢?

最开始研究这些问题的并不是计算机学家,而是社会学家和心理学家,他们的主要任务就是对社会进行抽象再抽象,从而找到在一定条件下普适的规律。在

社会学中有一个理论称为“结构洞”，即在整个人与人构成的网络中存在一些重要的节点，这些重要的节点具有一个特性：如果把这个人连同他所关联的圈子去除掉，整个网络结构会受到重要影响。基于此，计算机科学家们发明了影响力传播算法，其核心还是社会学中的理论。计算机领域存在许多影响力发现的算法及其变种，这里就不展开论述了。

行为时间序列分析

如果把一个人的所有行为从数据中抽取出来，我们可以绘制出一个人在网络上活动的时间分布图。一个典型的工作一族的行为时间序列是这样的：早上 7:00 起床查看早新闻，8:30 到公司上班开始浏览工作内容网页，中午 12:00 吃午饭，午休查看娱乐消息打发时间，下午 1:30 开始工作，傍晚 5:30 下班后开始休闲或是阅读，晚上 10:30 入睡前刷刷微博等。通过他在互联网上留下的痕迹以及发生频次在时间上的累计图，可以看到其行为频率随着时间的高低起伏变化，进一步可以应用前文提到的时间序列分析方法判断当时的场景，揣测用户的需求，预测接下来用户会有怎样的行为。

使用行为时间序列分析，微商一族们可以更加精准地掌握发帖时间，在合适的时间为自己朋友圈中的客户推送最准确的内容，并覆盖最多的用户群，获得最多的阅读量和关注度，从而达到促进销售的目标。如早间推送正能量鸡汤，午饭和午休时间推送产品，傍晚时分晒业绩，晚饭秀生活等。

行为模式挖掘

单一用户的行为也许具有随机性，但如果把一群人放在一起观察他们的行为，则具有很多的共性。特别是把一群相似的人放在一起的时候，这样的共性与规律性则更加明显。我们可以通过累计用户的数据，应用之前的关联规则算法把用户的相同行为模式提取出来。

行为模式提取有很多的应用场景，其中一个就是应用行为模式去发现相关人员，如违法或者是异常行为人员。除此之外还可以在教育行业供研究人员探究不同学生群体的行为差异，以达到行为干预与行为矫正的目的。

社团挖掘

社团是社交网络中一个重要的概念，它用于形容网络中联络紧密的一群人，

最早由计算机科学家 Newman 在 2002 年提出。社团挖掘就是从整个社交网络中找到哪些人相似并将其识别出来。通常借助聚类或者在社交网络图中找到几何结构的方式来进行。具体的算法已经超出了数据产品经理关心的范围，在此略过。

社团挖掘的用途多样。一方面，可以通过在社交网络中找到的社团判断人与人之间的关系，从而进一步针对其相似的喜好进行品牌传播和广告宣传，达到不错的效果。另一方面，可以找到有潜在危险行为的小社团，从而进行提前干预和遏制，以免造成严重社会问题。

传播路径分析

小学时候的我们或许都玩过这样的游戏，老师安排同学们列成一排，并告诉第一名同学一句话，让其依次告诉排在身后的同学，直至这句话传递到队尾的同学。有意思的是，队尾同学最终说出来的话往往和第一个同学从老师处接收到的信息大相径庭，而且队伍越长，信息就越失真。日常生活中，信息传播的例子不胜枚举。人们总是怀着好奇的心理来接收各种新鲜信息，俗话说“三人成虎”，说得人多了，谎言也会被认为是事实。

可是信息究竟是什么时候出现谬误传播的呢，在社交网络中便可以追踪得到。以微博为例，一条信息的传播必然经过很多人转发。从信息的生成、接收到传播，信息链条中的所有人都是信息的经手者。但是这些经手者是不太一样的，准确说是不对等的，他们有的重要，有的次要，主要原因是他们在社团中的影响力不同。因此通过信息的传播路径可视化，就可以了解到信息是如何在网络中进行传播的。如果加上时间，则更加直观明了。可视化传播路径不仅可以帮助有关机构惩罚造势者，粉碎谣言，也有助于对传播过程中的大影响力结点进行监控与舆论引导。

引爆点分析

巴尔科姆·格拉德威尔（Malcolm Gladwell）是美国作家，其著作《异类》与《引爆点》都是书市的宠儿，创造了销售的奇迹。在《引爆点》这本书中，作者旨在向读者介绍流行背后的因素。

一般的信息传播一段时间就结束了，而那些能够吸引大家眼球的信息，则会持续发酵，最终爆发成一则群体性的新闻，大街小巷老少妇孺人尽皆知。

一则信息要想被引爆，在格拉德威尔看来，由个别人物法则、附着力因素以

及环境威力法则相互作用，拆解来看则需要满足四个定律：一是短定律，即内容要短易于传播；二是新定律，即不是陈旧的新闻；三是好友定律，即需要一个社群的协助；四是快衰定律，即没有任何新闻能够一直占据头条。满足这四个定律的新闻才有可能成为引爆点。如果我们采用可视化的方式来看它，则是指在以时间为横轴的图表中立刻形成一个凸起的尖峰，那个尖峰便是我们要关注的引爆点。

上面从个人的角度（行为事件序列分析、行为模式挖掘）、社群的角度（影响力人物发现、社团挖掘）以及信息的角度（传播路径分析、引爆点分析）阐述了社交网络数据挖掘的方法以及应用，不知道对数据产品经理们是否有所启示。

9.14 排序：简约而不简单的事

讨论完文本与社交网络，让我们来看一个很基础的问题——排序。之所以说排序是一个老生常谈的话题，是因为大多数人在工作中都会遇到。工作业绩需要排序，找到重点客户需要排序，招投标也需要排序，看来排序无处不在。排序是节省社会资源的一种有效方式，一方面通过排序可以避免社会不必要的争论，能够使规则透明与公平；另一方面，排序有利于社会资源的集中分配和有效利用，毕竟排在前面的人使用资源的效率也相对较高。

本节中我们大致会讨论两个关于排序的问题，一是排序的规则方法，二是排序的操作机理。当我们在问规则方法是什么时，实际上是在问按什么指标进行排序，以及这些指标如何获得；而当我们讨论排序的操作机理时，实际上是在问当指标确定后我们是如何对待排序对象进行操作使其有序的。下面分别介绍。

9.14.1 排序的规则方法

我们从念小学开始就会数数，也自然会排序。3比2大，2比1大，这已经成为我们的一种自然认知，其过程明显就是找一个指标，然后按照从大到小或者从小到大的顺序排列。

看似好像很有道理，可是问题在于，这个指标应该怎么得来呢？大致有两种方法。

一种是指标融合的方式。针对业务罗列一些指标，譬如身高、体重等，每个指标其实都有各自的数值。然后给每个指标一定的权重，把它们融合到一起，得到一个综合指标。这种方式被称为加权融合。

另一种方式相对来说与上述计算方法不同，是一种基于网络的指标构造算法。可以尝试这么理解，以社交网络中的用户为例，我们要对这些个体进行优劣排序，实际上就是把好的找出来，优胜劣汰而已。社会上优秀的人一般具备两大特性，首先他们善于向优秀的人学习，其次有同样优秀的人认可与承认他们很优秀。以微博为例，善于向优秀的人学习就是成为他的粉丝，你关注的人是什么水平，也在一定程度上决定了你的水平；另外有优秀的人愿意关注你，也说明你很优秀。很多情况下我们会关注很多名人，但是这并不能说明我们优秀，只不过是满足了其中一条而已，如果我们想要迅速在人群中脱颖而出，就需要满足第二条，只有这样才能成为一个优秀的人。

道理说完了，那么究竟什么样的网络算法才可以解决这样的问题呢？这里介绍两个，一个是 HITS 算法，另外一个就是大名鼎鼎的谷歌公司所发明的 PageRank 算法。

HITS 网页无非一出一入，出即表示网站中含有其他网页的超级链接，入则表示有别的网页添加这个网页的链接。根据上述的讨论，一个链向众多优秀站点的网页，其本身应该也不错。另外，一个被很多优秀站点收录的网页也一定不错。对于 HITS 来说，从一个初始值开始，为每一个网页执行这样的指令，也就是说，出等于其指向的所有网页的入相加，入等于所有指向它的网页的出相加。如此往复迭代，就可以获得每个网页最终的出和入的数值，这两个数值就可以用来作为表征这个网页好与坏的指标了。

PageRank 算法相对来说对数学的要求更高，PageRank 的一个假设前提是，一个网页若被访问，只可能出于两种情况，要么是在导航栏通过键入 URL 直接进入，要么由别的网页链接而来，所有网页被访问的概率（即重要性）均等。在上述假设的基础上，通过循环往复的迭代计算，即可获得每个网页的重要程度值，这个值就是我们获得的指标。

对于数据产品经理来说，只要知道有这样两种方法即可。

9.14.2 排序的操作机理

不知道你想过没有，排序内在的机理究竟是什么？如果能够弄清楚这个过程，也许计算机就可以帮助我们进行排序了。

下面介绍的四种排序方式是计算机学科中比较典型的四类算法，在我参与的校招面试中也有许多未加准备的计算机专业学生面对这四类算法面露难色。此处纯粹是为了扩展数据产品经理的视野，并非要求大家掌握，毕竟这些内容不是数据产品经理的必修课。

冒泡排序

冒泡排序可以用八个字概括，即“比较相邻，交换顺序”。假设一系列杂乱无章的数字摆在面前，我们需要从大到小排列，则可以每一次都从中找到一个最大的数，然后在剩下的数中接着找最大的。由于一遍一遍找最大的数字，犹如液体中泡泡越大获得的浮力越大上浮也越快一样，故而称为冒泡排序。

插入排序

插入排序的原理有些类似于抓扑克牌的过程。没有抓牌的时候手上空无一物，当从多张牌中抓一张后，这个时候手上只有一张牌，因此必然是排好序的。接下来再抓一张，然后和手上已经有的牌进行比较并为其找到合适的位置。在此过程中只要保证，插入牌之后手中的牌依然有序就可以。因此，只要进行有限次，可以保证所有的牌都有序，即所有的数字已经排好了顺序。

归并排序

归并排序同样可以使用八个字概括，即“分而治之，分久必合”。分而治之是之前介绍过的一种思想，即把问题分解成规模更小但情形一样的子问题。例如，有8个数字需要进行排序，我们就可以拆解为两个问题，即从中间划分开，对两组数字进行排序，每组数字中有4个数。这样一直分解，直到问题变成最简单的一组中只有1个数（已经有序）或者2个数（若无序，交换即可）。而“分久必合”又由何而来呢？刚才的过程是分解，接下来就要进行合并。当两组有序的序列放在一起排序时，只要逐个比较两组数字，从头到尾扫描就可以将它们组成一个完整的有序序列了。这个过程是将两个有序的数列归整合并到一起，故而称为归并排序。

快速排序

同样用八个字概括快速排序，即“分而治之，向我看齐”。分而治之不用再赘述了，随机从所有数字中选出 1 个，把它调整到序列的中间位置，然后把所有比它大的数字放到右边，所有比它小的数字放到左边，就好比所有的数字都在向这个数看齐。这样不停地在左右两边分解问题并进行相同的动作，就可以得到最终的有序序列。

原来这么简单的排序问题还有如此多的神奇解法，其实排序的方法远不止于此，还期待着你去探索。

9.15 推荐系统：“今日头条”背后的秘密

“今日头条”是当下内容分发类应用中的明星，相关机构的市场调研数据显示，今日头条的用户每天在头条 App 上会花费超过 60 分钟的时间。这个现象之所以会发生，是因为头条 App 会根据你看过的文章和视频为你持续推送相似的内容，于是用户陷入了一种沉溺于自己喜欢的内容而无法自拔的情形。

推荐系统对我们来说并不陌生。我们去商场购物，店员会推荐本季的服装。我们去餐厅就餐，侍应会推荐热门的菜品。我们希望扩展人脉，好友会引荐其他人。这些推荐的过程是由活生生的人完成的，而在互联网上则是由计算机完成的。计算机根据收集的用户信息为用户推送特定对象（好友、资讯等）的计算机算法与工程化系统，这便称为推荐系统。本节会介绍一些推荐系统的方法以及典型思路，具体实现算法不做详述。

推荐系统可以无处不在，而其之所以能实现推荐，是由于它分析了你在网络上留下的痕迹，知晓了你的喜好。如图 9-8 所示，在一个典型的位置社交网络中，我们大致会留下这样一些信息：首先是人口属性以及与我们连接的好友；其次是一些内容文本。这些内容可以是新闻或者类似于大众点评上的商家信息，有商家必然存在位置信息以及与商家的互动，这些互动包括签到、对商家的点评以及对商家的评分。高校研究机构，以及以推荐算法为核心技术的企业中的达人们，正是基于这些信息不停琢磨精益求精的算法，以期效果的提升。针对每一种信息，我们都可以发明出奏效的算法，把不同的算法混合到一起又可以形成新的融合性

算法，下面就来介绍这些算法。

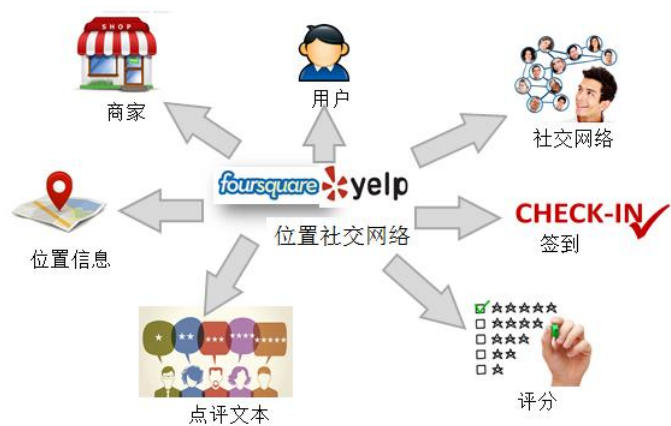


图 9-8 社交网络中的数据

基于内容的方法

内容这个词在这里得稍微变换一下，换成标签更加合适。我们使用携程或者美团点评时，都会按照一定的标签来筛选，标签可以说是店铺或者物品的标志。那么被推荐人呢？我们也会为其打一些标签，譬如学生、四川人等。基于内容的方法便是将人的标签和物品标签进行比照匹配或者相似匹配的方法。例如学生一般没有钱，因而匹配便宜；四川人可能喜辣因而匹配火锅等。如果遇到标签正好完全吻合那自然最好。

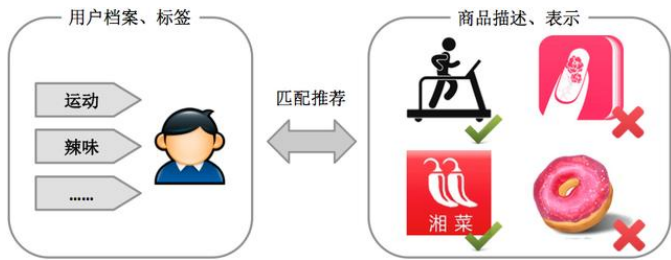


图 9-9 基于内容的推荐系统

协同过滤

协同过滤包含两种类型，一种是基于用户的协同过滤，另一种是基于商品的协同过滤。协同过滤可以分开理解。协同就是指需要大家的帮助，过滤就是通过

大家的帮助从海量的对象中筛选出中意的内容。那么基于用户的协同过滤就很容易解释了，一言以蔽之，在人群中找到和你相似的人，他们喜欢的内容或许你也喜欢，系统把这些内容推荐给你就是基于用户的协同过滤。同理，基于商品的协同过滤是把与你曾经喜欢的内容相似的内容找出来，认为你有可能也喜欢，然后推荐给你。

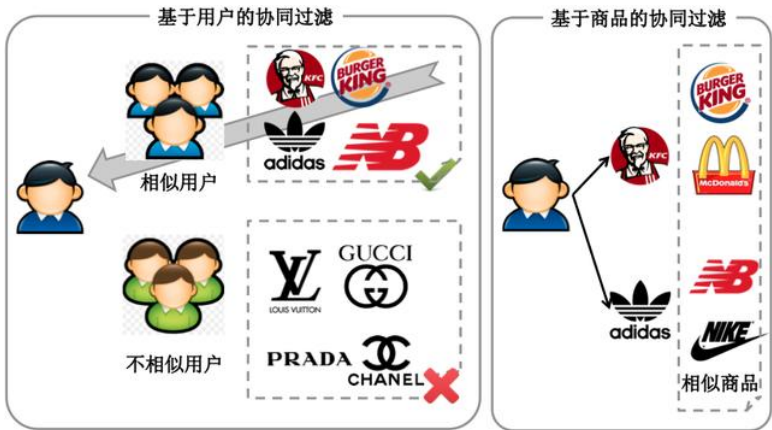


图 9-10 协同过滤的推荐系统

基于社交网络的方法

使用社交网络来进行推荐是一种源于社会学的方法，原理是社会学中的“趋同性”（Homophily），即一个社团中的人应该具有相似的兴趣。这个假设往往很贴合实际情况，我们很多时候进行购物是听从了朋友的建议。无论是单位，还是朋友圈，或是粉丝会，都是自然存在的团体。不同于基于用户的协同过滤寻找相似用户的过程，我们只需要在社团中寻找一些有话语权的代表人物，并把代表人物推荐的内容拿来推荐给别人就可以了。

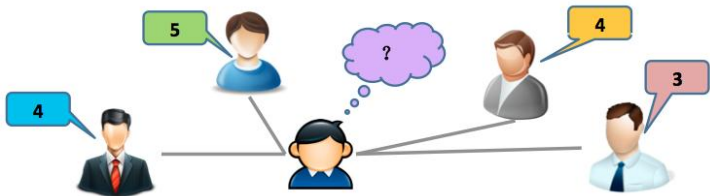


图 9-11 基于社交网络的推荐系统

基于位置的方法

我们的行为除了受到好友的影响，也会受到地理位置的限制。例如，尽管洛杉矶的某个商店在打折且商品质量不错，我们也很难逾越空间上的阻隔前往购物。因为在商品价格和质量因素之外，还有在逾越地理位置上需要付出的时间和金钱成本。

基于位置的方法源自地理学第一定律（Tolber's Law），即地理事物在空间上的分布存在着聚集性（Clustering）与规则性（Regularity）。例如我们喜欢逛的大商场或是城市综合体往往是商家聚集的地方，而以其为中心向外扩散，存在商家的可能性便会逐渐减弱，这也是为什么商家都希望将店铺开在城市中心位置。由此看来，推荐系统应该结合用户当前所在位置，并结合其他方法，为其推荐。

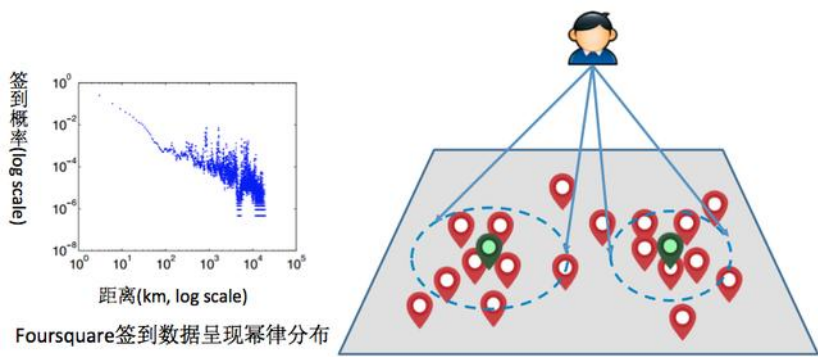


图 9-12 基于地理位置的推荐系统

基于点评内容的方法

用户的点评内容反映了用户的多种偏好，譬如口味、服务、价格以及位置等。一个商家的用户点评也决定了商家多方位的优势和劣势。例如，针对一位商务人士，其偏好为环境与服务，那么就應該为其推荐价格可能并没有优势的高档餐厅。如果为其选择性价比超高但环境嘈杂、服务冷漠的餐厅，可能会引起该商务人士的反感。这就表明商务人士更在意环境和服务，而对于价格因素没有那么看重。一个商家很难在多个因素上都具有优势，基于点评文本分析进行的匹配也算是一种折中后的投其所好吧。

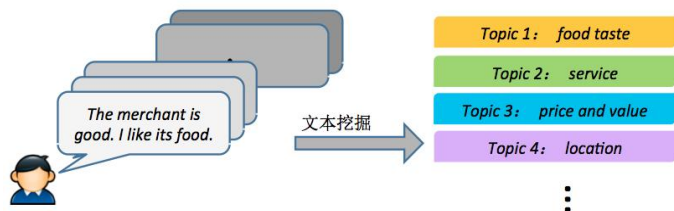


图 9-13 基于点评内容的推荐系统

基于时间的方法

用户的出行总是符合一定的规律，例如早晨总是在固定的时间上班，中午在固定的时间吃饭。尽管每个用户每天的作息時間不一定相同，但大致类似。结合每个人的行为习惯，就可以知道在什么时间应当推荐什么内容。譬如上班时间推送交通信息，临近午餐时间推荐餐厅信息等。结合时间情境，并考虑用户使用场景，在融合上述多种推荐算法的情况下进行推荐算法的开发，就是一种基于情境的推荐。

推荐系统也会遇到伦理问题。在《设计心理学2：如何管理复杂》一书中，设计学大师诺曼提到了一个概念——愿望线。人们不愿意走园林设计师们为他们设计好的路而偏偏选择从草坪上穿过，这条在草坪上由众人踩出的野路表达了人们从A点到B点的愿望路线，故而称为愿望线。在网络中，所有电子活动的痕迹就构成了人们的愿望线，这是对了解生活的有价值的补充。

发表论文的时候，你会引用他人著作中的内容。阅读书籍时，Kindle 会提供在喜欢的语句下方进行标注的功能。在电子商务网站上发现特别好的商品时会收藏或者购买它。这些痕迹都是社会性的语义符号。

推荐系统就是根据这些愿望线和语义符号而制造出来的。有了推荐系统，我们查找文献的速度加快了，这个人工智能助手为我们节省了时间。然而在另外一些时候，它也会变成“智障”。当在电子商务网站上浏览某件商品之后，那段时间所有访问过的网页上都会不停推荐相同的商品。还有一些时候，推荐系统让我们惊慌，你做了一件不想别人知道的事情，但是当朋友借用你的手机浏览新闻的时候，他们会知道你最近的偏好与一些不良嗜好。我们很难说推荐系统给我们带来了便捷还是徒增了烦恼。推荐系统诞生已经20年有余，然而在计算机伦理的道路上，它还有很长的路要走。

9.16 用户画像：隐私是个“伪命题”

上节介绍基于内容的推荐系统时，我们埋下了一个伏笔，我们谈到的内容实际上是标签，而其中人的标签构成了用户的画像。下面我们将对用户画像进行简要介绍。

最早并没有用户画像这样的提法，在此之前应该叫用户资料（user profile），主要内容是用户的个人基本信息，包括姓名、年龄、性别以及居住地等。

这些信息记录下来能有什么用呢？最先想到用处的必然是商人，他们善于利用蛛丝马迹找到商机。商机分为两种：一种针对潜在客户，即针对有可能在未来购买的人；另一种针对老客户，要争取让他经常来购买。但当时的商业相对来说较为原始，还处于人工筛选和人工判断的阶段，而且能够收集到的用户信息也有限，往往是通过一些问卷的形式来获取的，加之消费者天然排斥透露自己的隐私，因而商家能够获得的信息很少。

随着信息时代的到来，移动设备开始普及，用户越来越多的行为发生在线上而不是线下，因而我们有更加丰富的手段不通过询问消费者本人便可以采集到他们的行为，透过他们的行为又可以揣度他们的真实意图。伴随着用户行为的改变，商人的思想也与时俱进，他们把更多的事情交给机器去完成而不再依赖人工。在当下，能够被采集的信息也更加多元化，于是商人们创造了一个新的概念用来取代用户资料，这个新的名词就是“用户画像”。

说到用户画像，大家会在脑海中描绘怎么样的一个图像呢？是否像下图一样，一个人形身上贴上诸多标签。

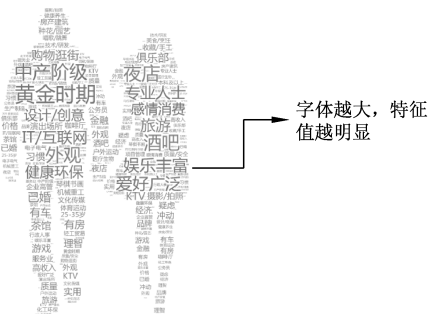


图 9-14 用户画像示意图

图片来源：http://www.sohu.com/a/115346825_461222

大家能够联想到这张图也不足为奇，这张图首先展示的是一个人形，代表了用户，而身上的标签表明了用户可能具有的潜在兴趣，即等于使用了一些描述性词语对用户进行描绘。可是这个画像做出来又有什么用呢？在商业的具体应用上，用户画像主要用于精准营销，而精准营销的使用场景多在广告，所以无论是以标签列表或自然语言的形式罗列每个用户的基本情况，还是像图中一样使用标签填充出一个人形，只要能够通过这些信息了解用户的兴趣，且便于在系统中检索，这个用户画像便是合格的。标签信息对于商业目标的意义在于言之有物，而便于检索则是对于描述同一件事的标签的标准化要求。

业界公认的用户画像大致分为四层，分别是原始层、事实层、模型层以及业务层，如图 9-15 所示。



图 9-15 用户画像的四个层次

原始层

顾名思义，用户行为的原始采集结果沉淀在原始层。例如使用某个 App 多少次，看了某篇文章多长时间，这些细节信息会被最大限度地完整保存下来。由于是最原始的信息，我们通过查看这些原始层的数据就能够还原用户当时使用设备的场景，因而它是高保真的。

事实层

事实层是对原始层数据在某个维度上的聚合。例如把一周以内或者一个月以

内的用户相同行为进行聚合，进而了解用户的经常性行为，并过滤掉一些随机和偶然发生的原始行为。原始层向模型层转变需要的是统计性的工作。

模型层

模型层是对事实层的抽象。例如事实层反映一个人一周内使用了5次携程应用进行酒店或机票预订，我们可以通过算法预测得到这个人可能是一位商旅人士。再例如一个人如果经常逛淘宝且使用支付宝进行付款，我们可以大致猜测这个人的消费能力。在模型层我们得到了诸如“高消费、夜店、财经”这些抽象意义的标签，并把它们贴到用户的身上，用以反映用户的兴趣和偏好。对于一位业务导向的同事或者一个外行人，他只需要查看模型层的结果就可以对该用户产生朦胧的认识了。如果需要更加清晰地理解细节，则可以借助事实层与原始层的数据。

业务层

业务层是对模型层的定制与组合。既然是业务，必然是针对具体的场景来进行的，假设在一次营销活动中，营销目标是找出可能购买某类西服的潜在人群，于是我们就通过模型层的标签进行组合筛选，并最终根据此次营销活动的目标分别为用户打上“会购买某品牌西服”与“不会购买某品牌西服”的标签。在这一过程中，我们需要从模型层的标签中，寻找具有商务标签且具有消费能力的男士进行广告投放。这就是业务层的意义。

最后以产品经理对于用户画像需要注意的内容来收束本节。

首先，标签分为静态与动态，即标签的时效性。有些标签（如性别）相对来说比较稳定，而有些标签（如一个人的消费能力）就较为波动。一个学生在毕业之前消费能力较弱，但是当他走入社会参加工作，消费能力就会改变。

其次，针对不同的业务，不一定需要四个层次都做，而且也不一定都能做。例如在教育领域，如果做业务是为了给用户提高成绩，那么更重要的就是给用户多贴一些知识点或者学习行为习惯的标签，而将其他标签的权值降下来。

最后，不同的行业进行用户画像的目的不同，但需要平衡好画像业务本行业的属性与广告行业的属性。仍以教育用户画像为例，如果实施的目的是为了拉新与促活，这本质上是一个营销问题，而非教育问题。而如果实施的目标是为了给校园管理者提供监管抓手，则更偏向于教育问题，而非营销问题。界定清楚问题的倾向，

有助于判断大公司内部业务的分工。

9.17 算法思想中的哲学内涵

数据挖掘的部分至此应该可以告一段落了，但总觉得意犹未尽。毕竟算法描述的是细枝末节的过程，而且仅仅可以用来解决一小类问题，跳出具体工作内容，究竟这些算法或者方法有什么意义呢？我们不妨可以将各类算法背后的思想抽离出来，探究一下这些算法思想的哲学内涵与对生活的指导意义。

蛮力法

蛮力法顾名思义，就是使用蛮力去搞定事情的一种方法。在工作中，有这样的情况出现，例如需要从网络中抓取 100 个对象的信息。许多工程师在遇到这样的问题时，首先想到的是编写爬虫以达到目的。这样做的出发点在于，认为程序更加快捷，而且更加通用，以后如果需要抓取 1 万甚至 100 万个对象的信息，也可以复用这样的程序代码。诚然，编写程序确实有这些优点，但那是对于最理想状况的解决方法。在项目时间和资源有限的情况下，或许手工解决会比编程和调试更加有效。解决问题，不是比技巧的高超，而是看效果的好坏与效率的高低。

在上面的例子中，手工收集信息是蛮力，使用计算机程序是技巧。编写计算机程序付出了相当大的固定时间和资源成本，只有在待解决问题规模增大的时候，低边际成本带来的价值才能够弥补固定成本的劣势。在计算机领域，很多事情是没有巧劲的，或者说花费了很多时间想出来的巧劲未必比使用蛮力做出来的要快。更何况很多的情况下我们未必能够在有限的时间内想出巧妙的办法，只能罗列枚举后一个个解决，一个个尝试。这暗示我们，在工作的时候，想不出方法不要怕麻烦，功劳没有就只能先沉下心来做一些称得上苦劳的事。没有事情是不麻烦的，使用机器学习算法训练数据之前还需要进行数据的标注和清洗，更何况数据产品规划这么复杂的事情。

分治法

分治法又称为分而治之，是一种把问题分解然后各个击破的方法，有些类似于《孙子兵法》中讲的“倍则分之”。在团队的任务管理和项目管理过程中，这种方法被广泛使用。任务需要被拆解到适合的规模，以使团队内的成员不至于工作

负荷不平衡。除此之外，拆解后的任务可以保证平行进行，方便统筹。

这里需要提到的是，分解问题有两种方式。一种是按照规模进行分解，一种是按照阶段进行分解。按照规模进行分解，是把大问题划分成小问题，但是小问题和大问题几乎是相似的，只是由于规模更小，因而可以更容易地找到解决方案，甚至不证自明。另一种是按照阶段进行划分的方法，划分之后的每个过程中需要解决的问题是不一样的，但是前一个问题的解决有助于后一个问题的解决，是首尾相连的。按照阶段进行问题分解有助于风险的分解管理，是一种有效的阶梯形式的行动方式。

贪心法

很多情况下需要进行规划，善于规划的人我们会仰之弥高，使用“眼光长远”等溢美之词来赞美他们；而对那些不善于规划的人则嗤之以鼻，形容他们“目光短浅”，只顾眼前利益。这种说法偏重情怀，从务实的角度来说，也会有这样一种可能，目光短浅的人能够呈现的短期利益是最大化的。我并非要否定长远规划，事实上长远规划与长远目标是两回事。树立长远的目标与愿景和阶段性的短期规划并不矛盾。

在计算机算法中，使用贪心算法是在求解问题时把当前的局部选择作为解决问题的最优选择，而不是从整体上加以考虑。在一些问题上，局部最优的组合就是全局最优；而在另一些问题上，局部最优的组合难以实现全局最优。能否实现全局最优取决于这样的问题是否具有无后效性，即某一个决定做出后，后面的状态不再受此前决策的影响，只受到当前决定的影响。

在技术风起云涌的时代，瞬息万变的市场环境不允许我们规划太远，即便进行长远规划，也应在过程中动态调整，抱残守缺注定要被淘汰。正如老话说的那样：“多鸟在林不如一鸟在手”，我们不要痴迷于长远的规划与设计，觉得有可能通过一种规划把所有的问题都解决，相对应地，比较实际的方式是遇到问题当下解决，日拱一卒地进行改善。

回溯法

回溯法是一种向源头追溯的方法，这从其英文翻译 backtracking 中就可以看出。在计算机领域大致可以类比一种搜索的算法——DFS（Deep First Search，深度优先

搜索)。我们之前介绍决策树的时候提过，搜索的过程仿佛是在一棵树木上找到一片你需要的树叶，你要顺着树干找到那片树叶所在的枝丫，然后继续“顺藤摸瓜”最终达成目标。但是我们又知道，从树干到树叶的层级相对较深，如果每次都从树干到树叶，当发现一条路走到头并不能找到树叶时，就需要向回溯源，然后沿着这条路的另外一个分支前进，循环往复。这种方式便是回溯法的来源。

回溯法带给我们的是一种试探性的策略，即错了不要紧，可以从错的地方回头，然后再次寻找一个新的路径去试探。创新很多时候是没有标准答案的，其过程中充满了试错（Trail and Error），回溯法给我们提供的正是这种创新方法在计算机上的应用。

计算机针对具体问题，有时和人一样也采用的是同一种尝试性的解决策略，即在大致选定了方向之后就向前进，直到遇到南墙才回头重试。但是回溯不能够成为懒得思考的借口，觉得从树干开始随机选择方向就可以了，这个想法是错误的。大多数情况下，细节问题可以边走边解决，但是在解决问题的道路上，缘木求鱼需要找对大路径，否则回溯的成本将增加，这也很难称为一种明智的策略。

分支定界法

分支定界法的英文是 branch and bound，其中包含了两层意思。一是 branch，意为分支，即将问题分解开。二是 bound，表示界定，即估出问题的下界以免无穷尽细分。这是一种不同于回溯法的搜索策略，我们可以很轻松地找到另外一种搜索算法和它对应起来，称之为 BFS（Broad First Search，广度优先遍历）。

BFS 适用场景大概如下，我们并不采取一条道走到头的方式，而是先针对摆在面前的各条大路都走两步，进行一下试探，然后再对某几条路进行筛选，再多走几步，逐步筛选。这种感觉仿佛是针对一件事先有一个思路和框架，然后再根据重要性谈细节。这也是麦肯锡的“金字塔原则”中必备的框架结构性的要求，切忌一下子钻到细节中。除此之外，bound 也提示我们要对每个分支中的下界，也就是最坏情况，有个估计。当 worse case 被预计到，并找到对策后，任何实际发生的情况便只会比这个情况好，而执行人也能够做到“泰山崩于前而我自岿然不动”了。

迭代法

迭代二字对于产品经理来说并不陌生，我们平常进行产品研发，难以一步到位

做大做全，而是要每次做一点，根据反馈，下次再完成一些，如此循环往复，这个过程就称为迭代。

《精益创业》中提倡的“精益”思维便是这个思想在创业中的应用。这样的过程能够给我们的启示不言而喻，凡事不是一蹴而就的，在工作中追求完美是一种比较耗时耗力的要求，往往受限于外界环境。为此，很多互联网公司打出了“完成比完美更重要”（Done is better than perfect）的口号。因而对于实际的工作人员来说，了解当下应该完成什么，分清楚轻重缓急，这一点十分重要。这是一种“要事优先”的思维，也是一种先搭框架后填内容的行动方式。

关于判断什么事情比较重要，很多人似乎比较困惑，往往感觉很无助，这里介绍两种思考方式帮助大家进行判断。一种是李开复老师介绍的“见诸报端”法，其原理大概是，假设每件事情都做砸了，后果都会被刊登在明天的报纸上，那么哪件事情的刊登会让你觉得最羞愧，那件事情就是对你来说最为重要的事情。另外一种方法称为“淘汰排序法”，其大致原理是，把事情按重要程度从高到低排序似乎很痛苦，但是从诸多事情中选择重要程度低的似乎要轻松一些，不妨逆向思维，一个个选出最不重要的事情，然后最后留下来的就是最为重要的事情了。

算法的思想源于哲学，但又在信息时代发扬光大，并最终在生活中得到应用。我们追随前辈的脚步，面对生活中的事情，多想一些其背后的机理，做一个有“理论偏好”的人。

第 10 章

面向产品经理的数据可视化

- 10.1 别人家的可视化：阳春白雪..... 222
- 10.2 工作中的可视化：下里巴人..... 227
- 10.3 用可视化“说谎” 230
 - 10.3.1 数据的误导 230
 - 10.3.2 逻辑的谬误 234
- 10.4 准备一份数据报告 238

总算从数据挖掘的泥沼中摆脱出来了，之所以有这样的形容，是因为数据挖掘部分是对技术与思维要求较高的环节。对于大多数产品经理来说，数学是令人头疼的事情，即便里面仅仅涉及加减乘除，但若密度过大，也会令人身陷囹圄，难以自拔。别怕，本章我们将转向一个较轻松的话题——可视化。

在这个部分介绍可视化是顺理成章的，按照前面文章中提到的 CRISP-DM 数据挖掘流程，最后一个过程便是实施，而对于数据类项目的实施，最重要的一个环节便是展现数据。本章中，你大致可以了解关于可视化的几乎所有内容，从精彩的图表和可视化案例，到可视化工具，再到使用可视化的一些特殊技能，最后我们会以可视化的某种正式形式，即数据报告，作为结束。

10.1 别人家的可视化：阳春白雪

2016 年 11 月，我在美国西海岸访问，曾在 Facebook（脸书）总部拍摄了一张用于展示其发展历程的图片。在图上我们大致可以看到一条波动的线条，线条上的每一个格子代表了 Facebook 历史上的某个年份发生的重大事件，伴随周围跳动的圆圈则是公司旗下多种产品的 LOGO，其韵律感与视觉效果还算不错。



图 10-1 Facebook 总部介绍其历史的大屏幕

或许说到可视化与 Facebook，大家更加熟悉的应该是下面这个显示了其全球用户连接关系的可视化图。该图是 2010 年 12 月由 Facebook 的实习生 Paul Butler 使用公司全体用户数据的一部分子集做出来的。至今仍在 Facebook 会议室墙面的

屏幕上滚动展示。



图 10-2 Facebook 全球用户社交关系图

图片来源：<http://radar.oreilly.com/2011/01/visualization-facebook-friendships.html>

上述关于 Facebook 的两个图都是可视化的绝佳案例，概括来看，一个使用时间线来展示，一个使用网络来展示。除此之外，深色的背景给予视觉更多的深邃感，而明亮的地方则更加突出信息的内容。这也难怪许多用于演讲场景的胶片都是以深色为背景的，面向政府部门的大屏展示也多选用深蓝色调。驾驭这样的颜色还是需要一定功力的。

除了大公司外，很多生活中的例子也可以被可视化。许多人喜欢咖啡但却不一定分得清楚各种咖啡之间的区别。图 10-3 将各种咖啡的区别通过可视化的方式向人们清晰地展示，每一个杯子通过图形比例的方式注明了各式咖啡在咖啡、奶等原料上的成分与比例。



图 10-3 不同种类咖啡的可视化

图片来源：<https://www.fastcodesign.com/3047340/how-your-brain-understands-visual-language>

再举一个离大家生活比较近的例子，我们平常出行涉及的公交线路图，例如地铁图，就是一种可视化的亲民示例。图 10-4 至图 10-7 展示了世界上两个霸气十足的城市的地铁线路图，一个是英国首都伦敦，另外一个是中国的新兴大城市合肥。上面的四幅图可以这么看，其中的两幅图（a）是伦敦和合肥这两个城市地铁线路在实际地图上的分布，而另两幅图（b）则是对照图（a）做了一些艺术的处理。譬如采用相同的颜色表示相同的线路，都是横平竖直或者倾斜 45 度角等，这些都是可视化创作过程中的技巧和要领。

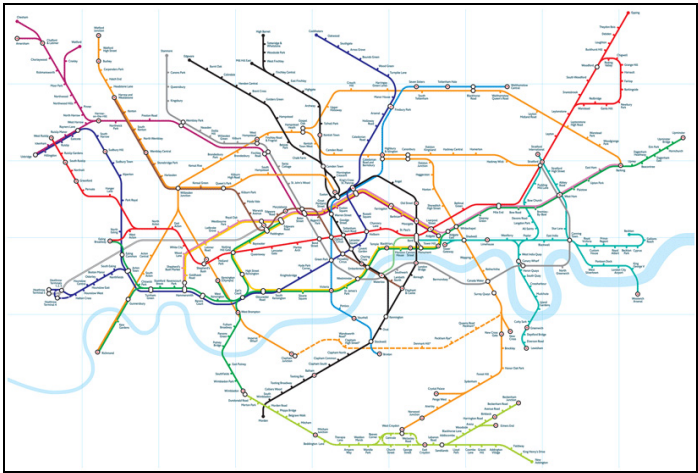


图 10-4 伦敦地铁图（a）



图 10-5 伦敦地铁图（b）

图片来源：<https://www.designboom.com/art/london-underground-map-reinterpreted/>

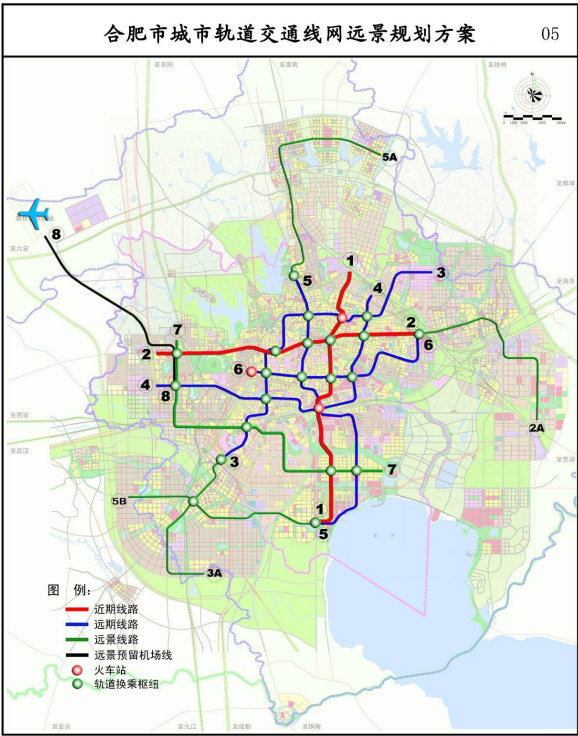


图 10-6 合肥市地铁规划图（a）

图片来源：http://imgbdb2.bendibao.com/hfbdb/20151/22/20150122114725_98051.jpg



图 10-7 合肥市地铁规划图（b）

图片来源：<http://hefei.auto.qq.com/a/20150703/011854.htm>

想了解更多的示例，可以通过输入 Infographics 进行检索，也可以通过多参观艺术展览来获得灵感。通过上面三个例子，或许我们可以毫不吝啬地说，可视化是科技与美的结合。一方面，图中展示的内容需要科技的手段进行测算与计算，而且需要使用科技的方式将其呈现出来；另一方面，色调的搭配、字体的大小样式无一不影响最终的呈现效果。因此使用“科技与美”来形容可视化一点也不为过。

前文已经介绍了历史中的可视化。下面让我们从信息传播的角度和人们接受信息与处理信息的角度再来审视一下可视化的必要性。有一项研究报告指出，人类阅读文字时，每分钟可以理解（注意，是理解，不是浏览）120 个字的信息，而如果变成阅读图片，每秒可以接收大约 8.96MB 的信息，可见信息通过图片被吸收的程度之大。目前也正发展使用 VR（虚拟现实）设备来辅助教学帮助提高单位时间内学生的学习效率，如若不然，VR 是不会被教育所接受的。

既然可视化需要结合“科技与美”，那么就需要满足一些基本的原则。清末启蒙思想家严复在《天演论》中讲到，文章的翻译要领在于使用了“信、达、雅”作为标准，在这里形容可视化的要领也恰如其分。

所谓信，表明图表传递信息是清晰的，没有遗漏也没有偏颇。所谓达，是指图表传递的信息能够直达用户，不会让用户困惑。所谓雅，就是对图表本身美学角度的要求，需要其制作是美观的、精致的。

满足这三点的数据可视化作品就算是及格了。那么一个可视化作品中到底有哪些信息可以传递呢？为了给一个图表添加上最丰富的信息，需要动用一切可以动用的想象力。我们甚至可以将五种维度的信息揉进一张图表中表示整体信息，例如性状（正方形、圆形等）、颜色（红黄蓝）、尺寸（大小）、位置（不同区域）、方向（动态变化）等。不过要想使可视化作品出类拔萃，没有艺术与人文的基因仅靠上述生硬的技巧也是难以维系的。有一句话叫“设计做得少，全因看得少”，因此多参观艺术馆、艺术展，留意生活中的小情趣，可以为可视化素材时时刻刻地加料。

可视化是“科技与美”的结合，真正优秀的图表是可以自己说话的。如果想再锦上添花，你只需要在向人们展示图表的同时配合一些介绍技巧即可。

如切披萨或蛋糕时的样子，每一块都代表了一个部分，各个部分构成了一个整体。矩形的代表有柱状图、条状图以及矩形树图。对于柱状图来说，将一根柱子切成类似竹竿的若干段，每一段代表了一个部分；条状图就是转置的柱状图；矩形树图则是在一个矩形中进行切分，分成许多个不同大小的矩形，每个矩形都代表了在整个部分中所占的比例。异形图有雷达图、南丁格尔图等。当上述几种构成类图形与时间相叠加时，就出现了堆积面积图与堆积柱状图。

比较

比较是在多个纬度上进行同等尺度的度量。与构成不同的是，构成只关心一个对象内部的各个部分，而比较则是比较多个对象。我们把表示一个对象的柱图和雷达图进行粒度上的扩展则可以用来对比不同的对象。除此之外，折线图也可以看成柱状图的抽象表达，若叠加时间，则可以用来表示趋势。

分布

分布是统计上的一个概念。在统计不同区间的样本数时，我们使用的是柱状图；当采样粒度足够细，各个柱体的顶点被连接时，就成为了折线图。这样的分布可以看成一维图像上的分布，而当这样的情况扩展到二维与三维层面时，就变成了热力图与三维图。

其他

其他的图表还包括用于表示不同对象间数据流动的关系图，称为桑基图。还有表示在不同步骤之间转化与过滤关系的图，称为漏斗图。

为了更加直观，让我们来通过一张图了解上述图表，大家可以在平时的工作中灵活多样地进行选择，如图 10-9 所示。

了解这些图，针对不同的业务选择合适的图表进行呈现，可以从业务的角度来理解图表。然而“正确选择图表”与“图表能够真正被做出来”还是有一定距离的。这段距离需要用图表中展示的数据来弥补。作为数据产品经理的你，应该把这样的事情交给谁？是研发还是前端？或者自己？下面提供了三种可以用来进行可视化开发的方式，供参考。



图 10-9 多种多样的可视化

商用 BI 软件

大数据行业中有很多公司围绕数据报表这一业务进行研发，大多产品形态以 SaaS 形式呈现。无须使用者掌握编程技巧，只需要导入数据便可快速上手，这是一种对大多数人都适用的可视化工具。运营人员与产品经理熟练掌握若干工具十分有必要，这类工具中的代表有 Excel 透视表、Tableau、神测数据、BDP 等。很多人在选择 BI 工具的时候，优先想到的是一些外部工具，殊不知用了许多年的微软 Excel 中具有透视表功能足已满足大部分的分析需求。

数据分析语言

对于数据分析师来说，掌握一门编程语言有多重要不言自明。拥有计算机专业背景的分析师会优先选择 Python 或者 R 这样的脚本语言；而具有统计学背景的分析师则更倾向于使用 SPSS 或者 MATLAB。对于产品经理来说，可以根据自己的需要，选择适合自己的编程语言。不用担心，现在的网上编程学院这么多，总有一款能够帮助你快速学习。在选择数据分析所用的编程语言时，还需要考虑另外两点，即团队整体的一致性与自身发展的前景。事实上任何编程语言都可以用来处理数据，只不过某些编程语言生而为数据服务，自带许多的功能组件供你调用，可以起到节省时间的作用。选择与团队一致的编程语言能够使自己在学习的过程中直接付诸实践，而选择热门且通用的编程语言则有助于自身未来职业的发展与个人的成长。

前端可视化插件

如果你的目标是通过网页发布数据分析的结果，那么前端人员的帮助就必不可少。这个过程需要掌握一些 JavaScript 基础，并非所有的产品经理都需要了解如何实现，但我也见过有产品经理具备自己编写原型的能力。在实现上述各种图表的前端插件中，最为通用的有 ECharts 与 Highcharts。

上述三种可视化开发方式都是从点子到产品的必由之路。虽然“条条大路通可视化”，但选择一条路坚定地走下去尤为重要，现在是你迈出第一步的时候了。

10.3 用可视化“说谎”

前两节说的都是可视化的美，正如玫瑰花也是有刺的，再美的事物也有可能误导我们。更何况有些人就是别有用心地使用可视化或逻辑谬误来诱导用户，以达到特定的商业目。

10.3.1 数据的误导

让我们来看一些使用数据“说谎”的例子。

第一个例子是“截断数轴”，如图 10-10 所示。在这个例子中，同样的数据被

分别展示在了左右两边。左边的柱状图数据相对来说上升陡峭，右边的柱状图相对平稳，两者的区别仅仅是因为左侧图 Y 轴的起点被修改了。很多情况下，别有用心的人会使用这种方式来向你展示业绩的上升势头，行业的强劲走势，但你要留心，看看他是否使用了这样的技巧。

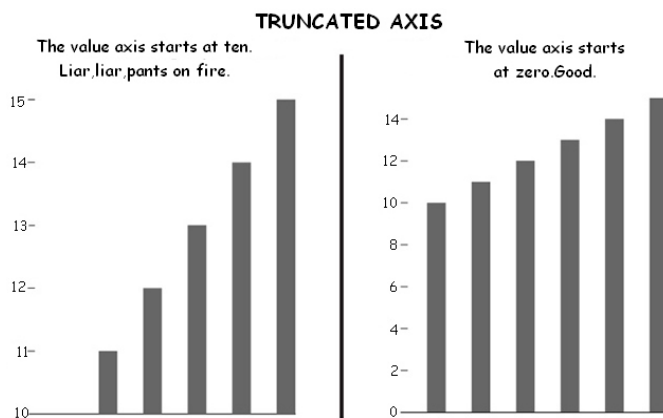


图 10-10 截断数轴

第二个例子是“双重数轴的假相关”，如图 10-11 所示。在这个例子中，两种不同的数据曲线被放在同一张图中。这张图拥有左右两个不同刻度的数轴，我们能够直观感受到的是，两条曲线此消彼长，因此容易产生一种误判，即两者是否有一些联系甚至正负相关性。实际上这个例子只不过用了假相关的技巧，把两个本来不相关但恰好有相似趋势的数据放在一起罢了。事实上，还可以将任意有相同趋势的数据进行组合，都可以得到类似的图像，但是并不能下结论说数据是相关的。

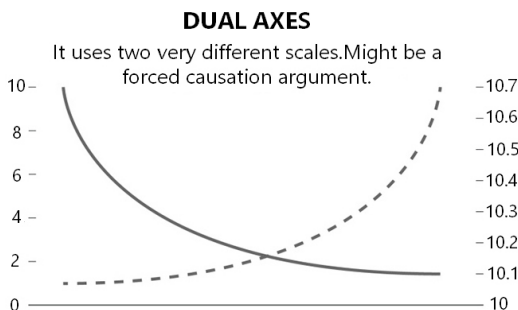


图 10-11 双重数轴的假相关

第三个例子是“绝对值与相对值”，如图 10-12 所示。当我们面对数字 100 的

时候，难以判断这是一个大数字还是一个小数字，因为缺少参照和具体的场景。如果是考试的场景，那么这就是一个大数；如果是一个销售指标的话，可能就是一个小数。同时，参照对象的选取也很关键。以考试场景为例，如果参照的对象都低于 100 分，那么 100 分就是一个大数；而如果参照的对象都考了 100 分以上，那么就很难说 100 分这个成绩是好是坏了。这个例子告诉我们，不要相信绝对值，而是要找到相对参照的值。

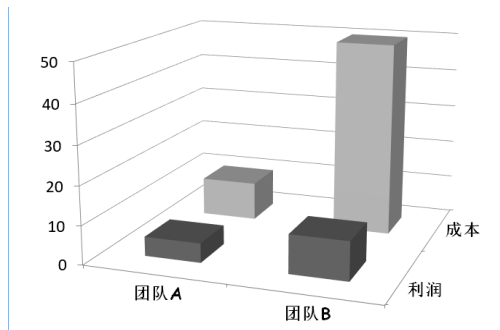


图 10-12 绝对值与相对值

第四个例子是“过分简化”，如图 10-13 所示。正如我们在前面介绍分类时提到的，我们可以使用决策树、逻辑回归以及贝叶斯算法把待分类对象一分为二，划分成两个大的类别。在计算机中，往往使用 0 与 1 分别标记这两类。但有时候，一些对象处于两个类别的中间。它们有可能属于 0，也有可能属于 1，只是概率不同而已。但迫于必须要将其划分到指定类中的限制条件，我们不得不做出 0 与 1 的判断。不得不说，这样的判别是武断和不准确的。这种使用数据一刀切的看待问题方式，在哲学上是一种对立的“二元论”世界观，而现实生活中的灰色地带使得“非黑即白”的论断站不住脚。

ODD CHOICE OF BINNING



图 10-13 过分简化

第五个例子是“用边长陈述事实，用面积带来幻想”，如图 10-14 所示。例如我们对数据 10 与数据 20 进行比较时，可以根据上节提到的图表展示法使用柱状

图来体现。但为了强调 20 比 10 不只好一倍的时候，就可以采用面积展示的方式。在进行面积展示时，将正方形的边长分别设置为 10 与 20，用正方体的面积来分别表示这两个对象，这样会令人产生一种错觉，即后者是前者的 4 倍，而事实是，后者仅为前者的 2 倍。

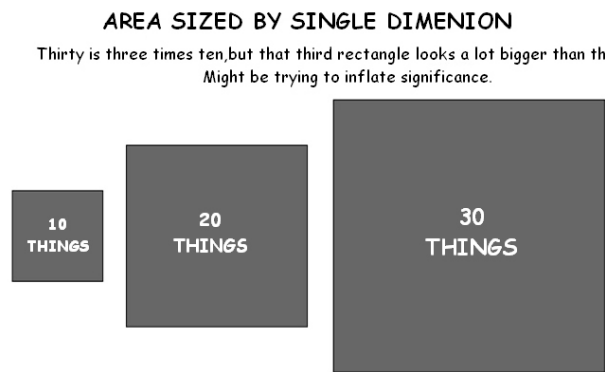


图 10-14 用边长陈述事实，用面积带来幻想

既然这些例子是可视化给我们造成的困扰，那么可能会有人提议不要使用可视化，还是原始数据靠谱，而接下来的例子又会让你大跌眼镜。

表 10-1 Anscomber 构造的四组数据

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

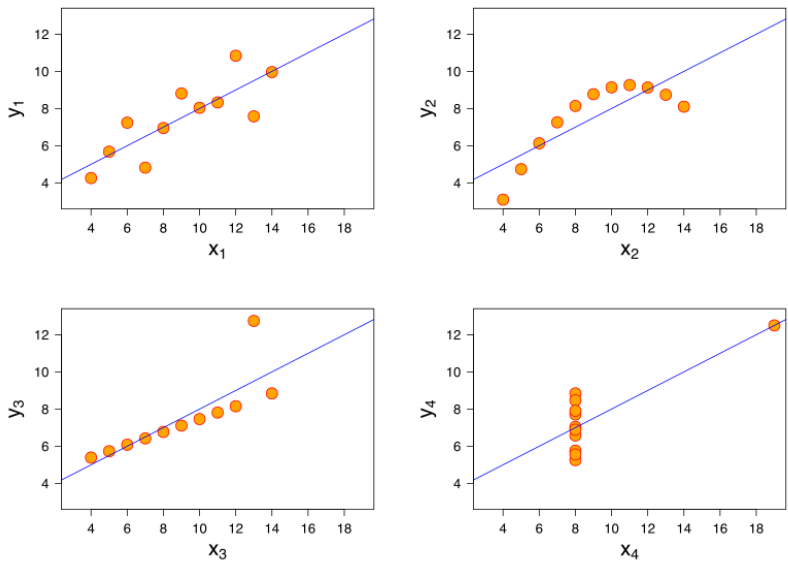


图 10-15 Anscombe 构造的四组数据对应的可视化

图片来源：https://en.wikipedia.org/wiki/Anscombe's_quartet

这个例子是统计学家 F·J·Anscombe 在 1973 年构造的四组数据，使用平均值、方差等统计量对数据进行简单的探索，结果发现，对于这四组数据来说， X 均值 = 9.0， Y 均值 = 7.5， X 方差 = 10.0， Y 方差 = 3.75，相关度 = 0.816，回归方程为 $Y=3+0.5X$ 。竟然完全一样！如果就此得出“这几组数据性质一样”的结论就错了。当我们将这些数据画成散点图时，就会发现这四组数据其实截然不同。看来，可视化和数据都有可能误导我们，需要实时警惕数据的误导。

10.3.2 逻辑的谬误

大致来说，可视化的“说谎”技巧可以分为两大类，一类称之为数据图表侧造假，另一类则是逻辑推理侧造假。

对于数据图表来说，采用的手段有造假、有条件挑选、平均与变形。具体来说，造假就是修改数据或对数据盲目四舍五入。有条件挑选就是选取最好的数据进行不实展现，例如第一天有 1 个人访问，第二天有 5 个人访问，于是得出网站

的访问增长率为400%，尽管这个数据没有问题，但是毫无意义，完全是好大喜功。对于平均来说，如一个人发了1篇文章，另一个人发了19篇文章，于是得出人均10篇的产出，这就好比以一个国家的人均收入来决定这个国家民众的幸福程度，是不太客观的，毕竟一个国家民众的幸福程度还和收入均衡程度有关。至于变形，上述的截断数轴例子就是印证。

对于逻辑推理来说，主要犯的错误是以偏概全与逻辑谬误。以偏概全的例子比比皆是，往往是由于滥用归纳法所致，才看到几个例子就轻易下结论。逻辑谬误又称包罗万象，大致可以从归纳法、演绎法、因果律与躲闪法等几方面展开阐述。

归纳法

所谓归纳法，就是提取共性的方法。这种方法的本质是使用一种化繁为简的方式来对待世界。众所周知，这个纷繁复杂的世界存在众多模式，人们发明了简化的机制来描述世界，以确保人与人之间传递信息的准确。从音素、音节与语标对于语言的简化，到数学公式对于自然科学的简化，无一不是简化的功劳。但由于简化，我们也会损失很多的信息，造成正确思维方法误用，从而形成一批归纳法的逻辑谬误。

首先要指出的是“质的量化”。这个世界上有感性和理性的认识，分别对应的是定性（quality）与定量（quantity）的衡量。如果非要给一个抽象的需要以定性分析的概念进行打分，就会使得原来有温度的抽象词汇变得冰冷从而出现了信息损失，虽然简化了评价的机制但确是一种隐性的逻辑谬误。为了破除这样的逻辑错误，只需要分清楚什么可以量化，什么不能量化即可。再浅显一点，抽象的概念几乎都难以被量化。

其次要说“浅层分析”谬误。这一类谬误产生的原因是“只看局部，不看全局”，犯了以偏概全的错误。现在大多数人说到周易与孙子兵法在商战中的应用都能信手拈来，甚至连刚刚入行的年轻人也能够引经据典。除了一部分人真的有学问外，另一部分人是用自己具有的简单分析能力在企图说服对方。而这样的分析处在浅层次，并不深入，在不明真相的群众眼中，这样的逻辑谬误是具有科技含量的。为了破除这种逻辑谬误，最好的方式就是深度思考，全面思考，谨慎下结论。

接着要提到的是“分类错误”。我们往往会用简单的分类模式来看世界，譬如

我们会依据行为事实将人分为好人、坏人；会根据所属关系把物品分为你的、我的；会按照地域不同将人们称为某个地方的人，并认为这个地方的人具有某一类显著的特点。一旦将某个人划入某一类，就等于给他贴上了标签。这是典型的先入为主的错误。要想驳斥这样的逻辑谬误，只需要将贴标签这种行为看成一个片面的并不科学的事情即可。

最后要说“举例论证”。能够被证伪的才能称之为科学，举例再多都不能够证实一件事情，因为一个反例即可推翻所有示例堆积构筑的理论大厦。所以破除这种逻辑错误的最好方式就是找到反例。

演绎法

演绎法就是将已知的真理或者事实进行套用和推广。在使用这样的思维方式时，容易过分扩大已知真理的受用范围而造成逻辑谬误。

首先说“引用权威”。这个社会存在很多的权威、专家、公知。大部分时候，这样一群人社会的精英阶层，不可否认他们的知识更为广博，信息渠道更加畅通，因而也更具威信。但是其所说的话一般是具备一定的场景条件和前提的。过分推而广之就会造成错误。这是心理学中的从众心理。

其次要说“引用传统”，也适用于上述的情况。这种逻辑谬误背后的逻辑是：早就这么做了，所以现在也应该这么做。无论是引用权威还是传统，只要我们能够找到这些权威和传统的适用场景和发生的上下文，并将当下的场景与之比较，指出不同，便可以说明其适用性的有限。

最后想说的逻辑谬误是“流行合理”。《乌合之众》一书算是对这类逻辑谬误的最好诠释了。人们跟风的习惯源自于对自我的不自信以及孑然一生、特立独行造成的安全感的缺失，因而人们更愿意随大流。但正如那句话说的一样：“真理往往掌握在少数人手里”，所以流行即合理也是一种典型的谬误。那有的人会说，看来真理真的掌握在少数人手里，这也是一种逻辑谬误。这个逻辑谬误既可以用引用权威的角度来驳斥，也可以利用接下来要说的“两难选择”逻辑谬误进行驳斥。

因果律

因果律是我们认识世界的一种简化方式，之所以将其独立于归纳法来看，是因为与因果律挂钩的逻辑谬误不在少数，以至于国家公务员考试也经常考察此类

错误。

我们先说一个最简单的道理，“出身因果论”。我们都听过美国种族歧视的故事，“小偷的儿子永远是小偷，法官的儿子永远代表正义”。这就是出身因果论的最典型谬误。而事实上在美国历史上发生的具体案例就是，小偷的儿子当上了法官，而法官的儿子却沦为阶下囚。这个世界上不乏咸鱼翻身的例子，我们最多只能说，出身好的人成为优秀的人概率更大，仅此而已。

接着是“先后因果论”。我经常举这样的例子：我辅导了小明，小明上了名牌大学。乍一听这两句话，还以为是我的辅导使小明考上了名牌大学，然而我的辅导发生在小明上小学的时候，你能说是因为我的辅导才使得小明上了名牌大学吗？这就是先后因果论的最好例证。为了破除这样的谬误，我们需要了解真实情况，看清楚因果的联系紧密性。

然后我们要提到的是“相关因果论”。真实世界中相关性远比因果性存在广泛，特别是在大数据时代，我们可以利用数据得出相关性关系，至于哪些相关性能够推断出因果关系就需要人肉眼去鉴别了。譬如我们通过收集数据，得出医院病床的数量和死亡率成正相关，于是我们推断减少病床数量能够降低死亡率，岂不是很荒谬？很多相关因果论比这个例子更为隐蔽，听起来很有道理，真正要说服自己还需要对多重原因进行逐一控制变量的分析。

再说一个“过分推广因果论”，网上又称为“滑坡谬误”。经常举的一个例子就是，如果我们允许同性婚姻，那么就会造成很多很恶劣的后果，譬如……这就是对因果的过分推广，本质上是上文提到的人们对未来不确信的安全感缺失，所以更多的时候喜欢未雨绸缪。

最后想说的是“本末倒置因果论”。它的意思是，我们忘记了因果的先后顺序，而把果当成了因，学习成功学的过程中最容易犯这一类错误。认为成功的N条法则只要做好了就能够成功，其实要先成功过才能够总结出这些法则，而非相反过程。

躲闪法

躲闪本质上并不属于一种思维方式，它只是一种“骆驼式地把头埋进沙中躲避风沙”的行为。当我们不知所措，或者无力反驳的时候，就会采取这样的行为

来赢得自保。

“诉诸感情”是首当其冲的，当别人向你哭诉，或者用道德来绑架你的时候，便产生了躲闪式的逻辑谬误，他并不给出具体原因，而是用强烈的感情来唤起共鸣，从而博得同情或者占据道德制高点。很多网上的虚假众筹或者极端左派用的就是这一招。

其次要说的是“混淆视听”。这就是我们常说的转移话题。具体表现有两种，一种是用与当下讨论话题完全不相干的信息和讨论主题来取代当下的论题，第二种是当找不到合适的过渡论题时就嘻嘻哈哈假装这件事情不重要从而躲避过去。对付这样的逻辑谬误，只要能够时刻提醒自己当下的论题是什么就可以了，不要被对方带到沟里。除此之外，“对人不对事”也算是混淆视听的一种，即用同一人的其他事情来取代对当下事情的讨论，或者可以看成“分类错误”的一个示例。

最后一个躲闪的例子是“保持中立”，看似两边都不得罪，两边的观点都不认同也不反对，但这样的行为实为一种逻辑谬误，因为它并不会对事情的解决有任何的帮助，只是在职场情商上可以为自己加分而已。

根据逻辑学的讲法，论点等于关键假设与论据的结合。所以质疑的途径也非常简单，要么指出关键假设的不成立与不普适性，要么就质疑论据的不客观性与偏颇。

了解数据说谎的机理，并不是要教会大家说谎，有的时候出于商业目的考量，你可能确实需要尝试使用一些技巧。但更重要的是，通过这些内容的学习，我们具备了一双慧眼去拨云见日，洞察是非！

10.4 准备一份数据报告

从数据的收集、汇聚管理，到统计、分析与挖掘，再到“科技与美”合体后的可视化，经历了九九八十一难，终于来到了数据加工的最后一个步骤。在这个步骤中，你要将所有的一切呈现出来，形成一份数据报告（其实很多 Web 数据产品就是动态可交互的报告），这里有一些技巧、原则或许可以分享。

报告的生产者

我们平常看到的数据报告大多来自图 10-16 中四种类型的企业。一种是行业内的第三方洞察机构，以 TalkingData 与易观国际为例，它们收集了一手的数据，并用此数据来对各个行业进行洞察。还有一类是 BAT 巨头的相应研究院以及第三方智库，它们更偏重市场、政策、社会与哲学层面，因而出品的报告也具有很强的可学习性。第三种是一些传统的咨询公司，因为咨询公司天生的强项就是做报告。最后一类是一些科技媒体，它们以其媒体的优势汇聚垂直领域的资料与数据，甚至以问卷的形式对一些问题进行深入分析。媒体人的笔锋有较强的个人情感倾向，因而可以作为辅助参考。



图 10-16 生产数据报告机构的分类

报告的分类

报告的类别大致分为 3 种，分别是行业趋势、用户研究以及专题分析。行业趋势与用户研究一个宏观一个微观，而专题分析却不同，有的时候是在垂直领域，有的时候又是针对某一问题，有的时候是针对社会热点的讨论与分析。行业趋势型报告中以一年一度由互联网女皇玛丽·米克尔 (Mary Meeker) 出品的互联网趋势报告，以及 Gartner 出品的魔力象限与技术曲线最为出名。用户研究型报告领域则是以拥有数据众多的阿里巴巴 (研究消费)、滴滴 (研究出行) 以及易观国际 (研究手机应用) 的出品质量最好，频率也最高。专题分析往往是在时代热点和官方政策发布后进行的，有一定的时效性。

数据报告很多情况下是免费的，也有一部分是收费的 (但似乎总可以找到渠

道免费获得)。为此,我们应该向曾经为行业贡献诸多报告的各个企业与个人表示敬意。



图 10-17 数据报告的分类

报告的制作

在具体实践的过程中,也需要经常性地来进行反思和总结,到底怎么样才能够写好一篇报告呢?提一些原则性的标准太过宽泛,说一些太过具体的又显得琐碎没有高度。我肤浅地将我们撰写数据报告的理论概括为 3P,即 Prepare(准备)、Present(展现)与 Polish(修饰)。

在 Prepare 阶段,最重要的是搭框架,即为整个报告的制作找到一个脉络和故事主线。这个脉络可以是平行结构,也可以是时序结构,或是递进结构,还可以是总分结构。无论脉络是上面哪一种,最好都能够为这个结构中的每个部分想出相同字符长度的主题,以求形式上的整齐划一。除此之外,在这个阶段还需要完成“概念界定、数据说明、研究内容、研究方法”等问题的确定,如果有需要,背景阐述(政策、市场、技术)等内容也可以在此阶段准备。

在 Present 阶段,我们要实现“逻辑清晰、结论凝练、图文一致、阐述明确”。应该达到两个基本效果。一是远观,即能看到主要结论和 slide 结构。对于整个数据报告来说,框架结构与脉络经过了 Prepare 阶段的整理,在这一阶段已经需要为其填充内容了。无论向数据报告的骨架中填充什么样的肌肉,都不能喧宾夺主,

导致骨架不再清晰。“眯眼测试”是一种常用来测试主要结论是否突出的方式。二是近看，即能了解详细信息与具体数据。在远观看清结构和主要结论之下，如果读者有时间则可以自行选择细节内容阅读，因此细节内容并非不重要，而是满足读者进一步的了解需要。第一部分的整体环节与第二部分的细节内容，任何一部分都不可以出差错。

到了 Polish 阶段，故事的渲染、适当的炫酷以及遵循文案撰写的原则又显得尤为重要。一些图表的复查工作可以在这个阶段开展，例如报告的基色选取（滴滴阿里的橙、艾瑞的草绿、易观的蓝），你需要选择适合公司传统颜色的方案以加强品牌的认可度。再如文字的图表展现，即如何将文字做成图，其中可采用的逻辑有层次、时序、递进、耦合等。又如全文故事的提炼，要能够找到一条主线贯穿其中，并为其起个名字，名字要短且好记，如“两学一做”“三严三实”等。最后检查图表五要素是否齐全，即标题、图例、刻度、脚注、来源，完备方算完工。

一切的一切都是为了言之有物。要做到尽善尽美实为难事，但我们至少可以端正态度，朝着这个方向一步步迈进，不求最好，但求更好。

第 11 章

向数据科学家再迈一步

- 11.1 能文：陪运营跟踪产品看效果 244
 - 11.1.1 传统运营的基本功..... 245
 - 11.1.2 数字化运营“三”话你知..... 248
- 11.2 能武：追研发把控进度出成果 251
 - 11.2.1 数据采集 251
 - 11.2.2 数据存储 254
 - 11.2.3 数据计算 256
 - 11.2.4 数据分析 258
- 11.3 能聊：跟随销售面向市场找思路..... 258

“三一学院”是英国剑桥大学中规模最为宏大、财力最为雄厚、名声最为响亮的学院，其主要教授社会与人文学科，以及自然科学学科，从其内容涵盖来看，本身就代表一种广博。按道理来说，一个学院应该聚焦，而为了培养复合型人才，其授课的内容横跨了人文和自然。其立意之高，自然使其毕业生也更为杰出。大家可能听说过牛顿是如何发现万有引力定律的，那颗砸中牛顿的苹果树就位于三一学院的前庭花园中。除牛顿外的知名校友还包括拜伦、罗素、尼赫鲁、查尔斯王储等。

在互联网领域有一种讲法，叫作互联网铁三角，指的是产品、研发与运营这三股力量，这三股力量共同构成了互联网迭代更新的基石。

之所以在这里提到三一学院的大名，大致有两个目的。一是互联网本身是一个大命题，正如人类社会一样，需要的不仅仅是某一个方面的专业人才，更重要的是培养将各个领域串联起来的能力。正如同现实中的三一学院将人文与自然进行串联一样，进入互联网或者科技圈的人士更需要广泛涉猎才能够适应职业的发展。另一方面，互联网铁三角的“三”正好印证了“三一学院”的“三”。“三位一体”中的“一”是产品，而“三”则包含了运营、研发以及销售。

以产品为核心，运营、研发、销售围绕周围，谓之数据产品人的“三一学院”。

11.1 能文：陪运营跟踪产品看效果

说起运营人员和产品人员的关系，知乎上有个形象的比喻，说是后妈与亲爹的关系。亲爹与后妈共同承担照看儿子的义务，但是亲爹却不喜欢后妈对自己的儿子（产品）指手画脚，也担心其没有能力、意愿与耐心养活自己的儿子。与此同时，后妈也很纠结，一方面得像亲妈一样展现自己母爱的光辉，另一方面也体会不到与儿子的亲密。

为什么会有运营这个职位呢？这个职位原本是与产品经理职位混合在一起的，但是由于产品工作的增多，分工越来越细，最终从产品中分离出来。运营工作伴随着互联网的发展而有所变化，但是万变不离其宗，其最主要的工作职责是进行产品推广，促进用户使用，提高用户认知。我们常说的“要会营销自己”就

是把自己当作了产品。依照这个思路，我们的社交过程就可以看成自我运营的过程。我们在职业上的履新、发展与选择就是把自己推广出去的过程。我们不停地与外界社交并且作为社会的一份子为社会提供自己的一技之长，企业作为买方会调用与使用这些能力。另外，总会有人提出一些新的观点去引领行业，我们称这些人为“网红”或者“公知”，他们不停地提升大众的认知，以获取市场更加广阔的空间。所以从这个层面来说，我们每天都在做运营。

无论是传统运营，还是大数据时代的数字化运营，都会对产品各个生命周期产生至关重要的影响。

11.1.1 传统运营的基本功

从传统运营的角度来说，运营主要包含内容运营、用户运营以及活动运营三个方面。

用户运营

用户运营的关注点自然是在人这个层面。对于某个产品的用户来说，其与产品的关系大致可以分为初识、相识、相知、相恋、相弃这几个阶段。就好像恋爱的周期一样，我们称刚才列举的用户使用产品的几个阶段为用户对特定产品的生命周期。在不同的周期，用户会呈现出完全不同的行为特征，就好比你在恋爱的不同时期会与对方做不同的事情一样。

每个产品人员应该都希望用户对自己的产品如痴如醉，迷恋极深吧？但用户是不可能全部都处在同样的状态的，根据用户的行为我们就可以将用户划分成多个类别，每个类别中的用户具有相似的行为特点。以去某个餐馆吃饭为例，我们把用户在这个餐馆消费的频率、最近一次消费的时间以及消费的额度放在一起就可以对所有的用户有个大致的划分，这样的模型称为 RFM 模型（Recency、Frequency 与 Monetary）。在数字化运营中还会提到。

当我们从中发现一群高端客户之后，我们就要把他们保护起来，经常给予一些关怀或者优惠，使得他们能够经常光顾，或是替产品与平台说说好话，做做广告。由于这些有价值的用户往往消费层次也不低，都是购买了某种高端商品或者服务的用户，于是在普通用户之上我们就需要打造一个能够象征身份和地位的用

户群，这个用户群称为会员。

说到底，用户的类别以及是否为会员并不是一定的，例如你可能为了看某部电影而购买了视频网站的会员服务，但是这并不代表你永远会是会员。当会员失效，你自动又回到了普通用户的圈子，特权瞬间消逝。

纵观用户与产品互动的若干个阶段，用户运营对应在每个阶段的工作其实不太相同，但其内在都需要进行“开源、节流与刺激”。开源是指推广产品，拉到新的用户，譬如在一些网站打广告，或者在群里发软文等。节流包含了对现有用户的留存以及对已经相弃用户的挽回，尽管用户由于一些原因可能并不喜欢我们的产品，但是仍然要找到他离开或者即将离开的原因，然后留住他。刺激则是针对转化与付费来说的，运营用户的目的还是要慢慢引导用户使其走向付费的道路。

在过程中需要注意的是分时机、分场合、分对象制定运营策略。

内容运营

当你看今日头条的时候，消费的是内容；听“吴晓波频道”的时候，消费的是内容；逛淘宝的时候，消费的是内容。在上述的媒体平台、自媒体平台或电商平台中，最为有价值的还是新闻、音视频以及商品等内容。

这些内容有什么样的特点呢？如果观察这些内容，会发现它们都可以用于消耗用户的时间或是金钱。能够延长用户在产品或者平台上的投入时间是一种好的现象，这说明用户被内容所吸引。对于金钱来说，值得花钱的东西自然是对我们有意義的东西，没有人会在没有意义的事情上花费金钱。

内容运营是伴随着用户运营而产生的，其主要的目的也是为了提高用户的黏性以及活跃度。当你发现一个网站的文章值得阅读或者某个公众号的每日推文很合你的口味时，你会愿意每天登录和查看，这就是用户对网站和公众号的黏性，通过登录和查看，其在产品或者平台上的活跃度也增高了。

内容是一个真实存在的事物，如果把产品比作一个厂房，那么内容就是这个厂房所生产的物件，产品的运营人员就是这个厂房中生产任务的承包者。之所以称他们为承包者，而不是工人，是因为有的时候他们自己产生内容，有的时候他们让其他人（譬如用户）来产生内容。且不管用户生成内容的过程，运营人员对于内容往往需要经历“采集、创造、编辑、审核、推荐、专题、组织、呈现”等

多道工序。

在这么多复杂的工序中，对能力的要求主要集中在文案与创意上。

如果我们以更高视角看待这类问题，可以把生产内容与存储内容以及消费内容看成一个“内容的生态链”，就好比企业中的供应链一样，获取原材料，再加工，变成成品再销售出去。对于一个关注内容质量或者依靠内容制胜的产品与平台来说，在这个链条的不同阶段需要做的事情却不太一样。对于原材料，我们需要的是采集、编辑与审核；对于再加工，需要的是创造、组织与呈现；而对于销售内容，则更多需要推荐和专题。

活动运营

曾经有一个这样的小故事，一个用户靠着不停体验一些新出品的互联网应用，靠给这些应用多提意见，多生产内容，并且参与互联网公司招揽用户的活动，赢得了许多礼品，进而在在线网站上售卖这些礼品赚了一笔钱。这些礼品小到记事本、台灯、U 盘，大到相机与电脑，无所不包。我曾经就因体验某创业公司的 App 而造成后台流量流失，后获得了该公司一个路由器作为补偿。为什么会给用户白送礼品呢？这实际上是在通过发礼品的形式讨好用户，以增加用户对网站的好感。一个产品或者网站举办这样的活动就称为活动运营。

很显然，这样的活动绝对可以对用户的黏性以及活跃度产生切实的促进作用。从另一个角度来看，由于用户的活跃度并不完全基于其对产品或者内容的忠诚，而是基于对活动奖品的忠诚，于是这样的忠诚度在活动结束之后大部分将会消失，所以我们可以说，这样的黏性和活跃度是短期内抬升的。自然，进行活动运营的同学很明白这一点，而且他们的目标就是在短期内抬升某项指标。当然，在这样的过程中如果能够通过活动使得用户获得离开活动后的黏性则更为妙哉。

活动分为独立活动与联合活动，独立活动就是指产品或者平台网站自己搞的活动，而联合活动就是产品联合其他产品搞的活动。其形式可以是一些团购打折、会员邀请、签到激励、线下活动等。

为了完成活动运营，运营同学可谓要费一番脑筋。从文案撰写到流程设计，从规则制定到成本预估，从预期收益到效果统计再到改进措施，整个流程走下来相信必然会收获满满。从这个角度看，这样的活动运营更像是一种策划，一种组

织，于是你在学校组织活动的的能力就可以派上用场了。看，是不是很神奇，原来以前的那些活动没有白白参加更没有白白组织，锻炼这样的能力还可以辅助工作。

11.1.2 数字化运营“三”话你知

数字化运营是商业数据驱动决策的最好例子，离不开数字与数据。

在我开始接触数字化运营的时候，我很好奇，难道业务不知道使用数据来驱动业务增长和企业发展吗？答案显然是否定的。业务人员早就知道用数据来驱动业务增长，使用 CRM 和 ERP 系统就是最好的证明。只不过那个时候移动互联网的浪潮还没有袭来，数据采集的触角还没有伸展到方方面面，自然数据的驱动能力就很有限了。而且以 Hadoop 与 NoSQL 为代表的分布式存储、计算等技术尚未成熟，因而数据的实时性无法保证，进而造成数据驱动的决策时效性受到影响。当然，对于全球的企业来说，如果大家都在同一个水平上，决策都略有延迟，整体上是看不出竞争差异的。这个时候如果一些企业优先使用技术打破这一僵局，并且在业绩上表现较好，就会赢得市场的追捧，行业也将争相效仿，这反过来也将促进技术的广泛传播。历史的经验是如此相似，军工技术推动战争，反过来战争也点燃军工类技术在世界范围内广泛传播。历史的车轮再滚动几十年，实体国战争似乎不再我们身边萦绕，但是商业战场的硝烟却愈加弥漫，大数据技术便是赢得这场战争的又一利器。

在数字化运营领域，有三个概念不可不知，分别是黄金公式、海盗指标以及客户关系管理。

黄金公式

对于一个店铺来说，收入是核心。任何可以帮助店铺提高收入的方法都是店铺追求的。店铺收入的黄金公式如下。

$$\text{收入} = \text{流量} \times \text{转化率} \times \text{客单价}$$

这个公式中的每个指标在数字化时代都可以迅速计算，也可以通过更多样的方式提升。

我们经常听说漏斗这个概念，其无非是在说流量与转化率的问题。其中流量

就是入口，标志了你能够在多大的范围内去推广你的产品以接触更多的用户，而转化率的则对应的是用户的黏性，即促进用户的使用，甚至提高用户的认知使他付费或者心甘情愿帮助你传播产品。既然是采用漏斗来看运营的成效，那必然要牵涉数据。对于商家来说，其关心的是每天有多少客人（流量）光顾，有多少人成交（转化率）以及每个人花了多少钱（客单价），这些都是漏斗中的数据，因而从数据的角度来看运营，可以将运营进行数据化与量化，而这样的量化最终将作为运营人员的考核标准。

海盗指标

海盗指标（Pirate Metrics）是一种 AARRR 的模型。AARRR 模型非常出名，是由硅谷的 PayPal“黑帮”成员 Dave McClure 提出的，指的分别是 Acquisition（获取）、Activation（激活）、Retention（留存）、Revenue（收益）、Referral（推荐）。

对于 AARRR 模型，有两种不同的理解。一种是漏斗式，一种是手段式，如图 11-1 所示。

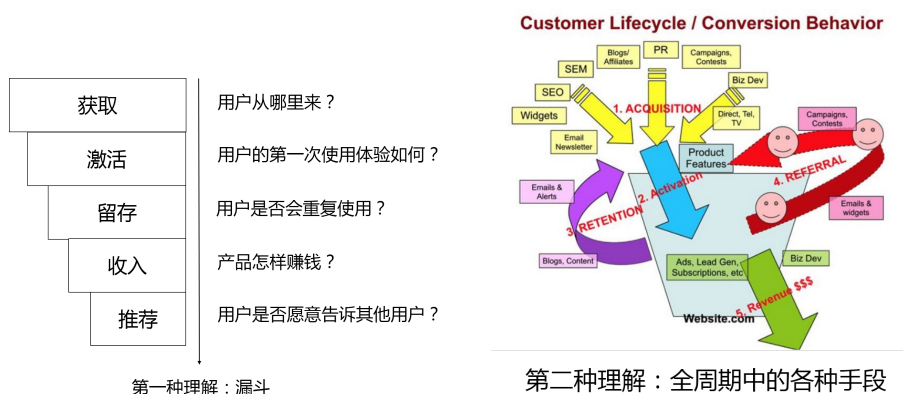


图 11-1 数字化运营中 AARRR 的两种理解

对于漏斗式，AARRR 代表的是获取、激活、留存、收入、推荐，分别表示五个不同的时期与阶段，且不同的阶段在时间上是先后关系，只有先获取用户，才能够让用户激活。这种思考方式是用一根用户参与产品由浅到深的主线贯穿起各个阶段的。对于手段式来说，AARRR 并非时间上的先后关系，而是同时并存的。例如对于获客来说，就是右图左上角部分列举出的多种方法，如 SEO、Web 浏览器广告插件，或者是 PR 等。再如，很多媒体类网站上都会集

成分享插件，这个插件就是用来引导用户分享推荐的，给用户提供一个便捷的手段去推荐产品。

客户关系管理

客户关系管理（CRM）是一个老生常谈的话题了，其中的 RFM 是谈论此类话题时必不可少的对象。R、F、M 分别是近度（Recency）、频度（Frequency）、额度（Monetary）的英文首字母，该模型的主要用途是对用户进行分群，进而针对不同的人群采取不同的策略，图 11-2 中左侧的两个子图就是示例。用机器学习的观点来看，RFM 是三个特征，对用户分群的本质是聚类。右边的两个图，是我们结合平时的业务对教育领域的学科资源与教育主体老师进行分群后，使用波士顿矩阵进行可视化的效果。

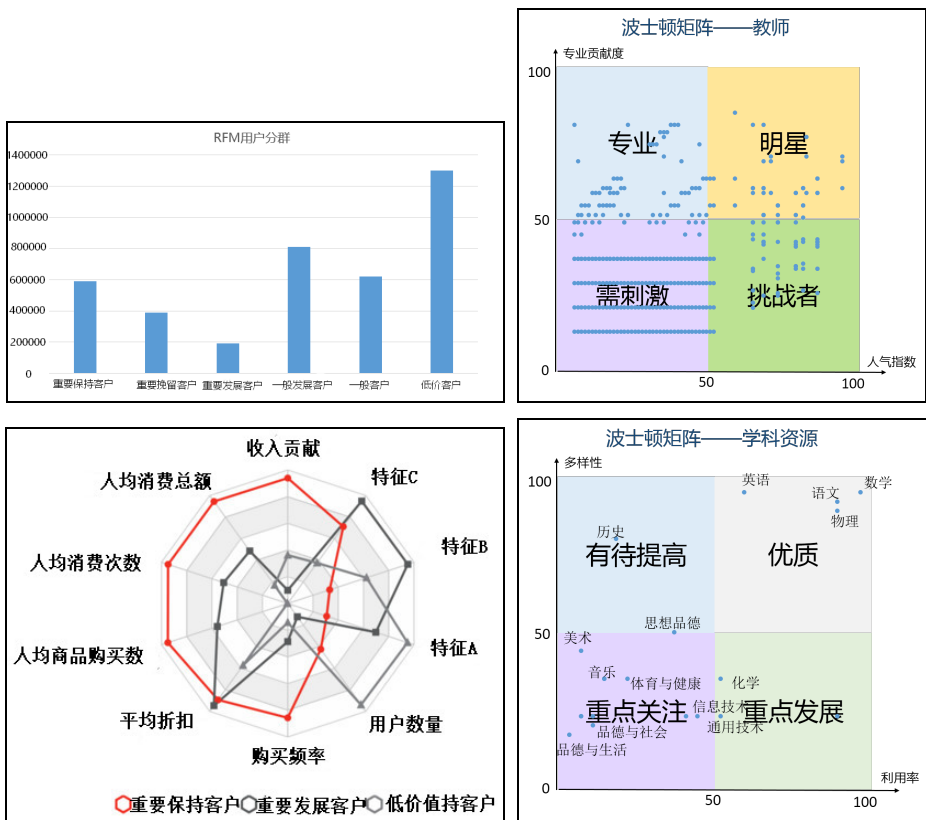


图 11-2 数字化运营中的 RFM 模型

对于运营的另外两个兄弟来说，产品与运营息息相关，反过来运营经常与客户打交道，也可以收集许多用户的需求反过来改进产品。对于研发来说，某种程度上是在帮助运营打造系统，例如开发一个抽奖的程序，都是需要研发进行投入的。运营在互联网铁三角中是冲在最前面的，其次是产品，最后是研发，于是运营人员的呼声往往是不可忽略的，千万别把运营人员当成了客服。

11.2 能武：追研发把控进度出成果

既然产品、研发与运营是互联网的三板斧，那么就不得不说说大数据时代的研发了。我一直在困惑，究竟应该如何向广大的产品经理们讲清楚什么是大数据的研发。说详细了，必然使其困惑，而且才疏学浅的我也不可能讲得面面俱到。说粗略了，产品经理又会觉得不痛不痒，毫无感觉，产生不了实际效果。再加上大数据的研发本就包罗万象，门派众多，云计算、虚拟化、微服务都算是与大数据紧密相关的，那么到底什么该讲，什么不该讲呢？

我想和各位产品经理同学讲一讲大数据的代表性工作——Hadoop 平台以及围绕该平台大致需要做的一些事情。我们并不会具体到某一种语言，也不会具体到某种技术，而是从宏观的视角，把产品经理在数据上扎根生出的藤条蔓延向原本只有研发工程师们才能够触及的墙角与远方。并在每一处需要精心雕琢与深究的地方稍加缠绕，略微展开谈一谈这些宏观过程背后的机理。

还记得在介绍数据挖掘相关技能的时候，我们借助 CRISP-DM 流程阐明了数据挖掘与数据分析的标准化步骤。如果将那个时候的介绍看成逻辑层面的流程，那么这里我们想沿着物理层面的主线来向各位数据产品经理展现什么是大数据的研发。具体来说，我们会按照数据采集、数据存储、数据计算以及数据分析这样四个步骤来介绍。

11.2.1 数据采集

现如今，相信很多人手上的手机已经变成了全屏幕的触屏手机，而如果时光回到十年前，你拿起那个屏幕尺寸更小，手机功能更少的键盘机，不知会作何感

想。那个时候的手机对于我们来说，只有三个意义：电话、短信、小游戏。如果你的手机可以播放视频，那一定是诺基亚与摩托罗拉的高端机型了。

在那个时候，我们通过移动电话设备产生的数据无非也就是通话与短信数据，这些数据由三大运营商管理，它们知道我们在什么时间给谁发短信或者打电话，甚至还可以知道我们发了什么，说了什么。那个时候的数据采集量小到可怜，而且大多数还握在运营商手中。

随着屏幕尺寸的扩大，Android 与 iOS 系统的入局，整个手机生态重组。人们还是习惯于看大屏幕的手机，习惯于使用一根手指去戳屏幕，习惯于在自己的电子设备上装上各色花花绿绿的应用，以至于在各种公共场合都低头猛戳手机，甚至还收获了一个专有的名词“低头族”。

这个时候出现了微信、微博，人们开始习惯使用微信给好友递送消息，而不再使用短信。这样的现象也改造着运营商的业务，以前每个月最需要在意的是套餐中包含多少条短信，而现如今则最在意每个月的套餐中流量有多少。尽管运营商还是掌握着海量的数据，但是能够收集数据的人再也不止运营商了。手机的生产者需要为手机出厂生产操作系统，因而成为了天然的数据收集方。除了手机生产者，其中各类 App 商家也能收集数据。这样，数据进一步被积累，除了以前的运营商（地皮所有者）之外，手机操作系统厂商（房屋建造者）与手机应用开发者（家居厂商）都成为了数据的拥有者。

原本数据只是一口池塘，池塘中仅有的几条大鱼难以翻身，而现如今风云突变，池塘变成了湖泊，不仅这几条大鱼能够鱼翔浅底，而且随着湖泊生态的演变，鱼群变得越来越密集，每条鱼都可以自由吮吸着湖泊中的融氧，变强，变大。就在此时此刻，湖泊的平面再一次的上升，而鱼群们也更加跃跃欲试，这个湖就是我们现在知道的大数据。

那么数据是如何被采集的呢？数据的采集方式大致可以分为两类，一类谓之传送带式，一类谓之渣土车式。且听我细细为你道来。

送带式的数据采集

采集的前沿阵地和数据存储的大后方之间建立了一个传送带。一旦有星星点点的数据被采集就会被立即送上传送带，经过一段距离的传输就会被递送到可存

储的地方存储起来，或者是被计算。就好像很多煤矿或者石料企业现场的一条长长的传送带，前方的挖煤机或采石机但凡能够搅下半点原料，这些原料就会被立刻上传送带，送到后方。尽管在实际的场景中，这些传送带的后方有可能也是一些仓库，但不妨把这些仓库看成在数据采集档口上的第一道存储服务器。在这样的情况下，运送渣土或者数据的通道是一直建立的，存储渣土和数据的仓库也是时时刻刻在运作的。只要采集数据的过程不停止，那么运行这个传送带的机器与看守仓库的人就不可以下班。

在这样的实时数据采集过程中，人的注意力不可以放松，需要一直专注于整个流程，我们称这种方式为长链接，就好比人的思想之弦一直紧绷，一刻不松懈地关注着另一端很长时间。与这种方式类似的数据处理方式称之为数据流处理，打电话就是一种典型的数据流处理形式，在整个打电话的过程中，听话人需要一直关注对方在说什么，需要通过听辨声音来判断通话是否还在线。

渣土车式的数据采集

在这种方式下，往往是挖掘机把挖掘到的土（数据）放到渣土车上，但是这些车不会立刻离开，而是要等车装满后再听从现场指挥的命令按照次序离开，并将土送到指定的地点，这样的方式因为需要将数据积累到一定的量，因而是一种相较前一种数据采集方式更加随意的方式，它的随意就体现在数据并不要求实时，而是一段一段时间发送一次。

因此，在挖掘机工作的时候，土方车的驾驶员（传送人）与仓库的保管员（存储人）可以稍加休息或是处理其他事情。这种数据采集的模式称之为短链接，顾名思义，人们并不需要死死地盯住一件事情，仅仅需要在需要处理的时候响应它，在处理完之后转而去做别的事情或者处理别人的请求即可。目前大多数 App 内的数据采集使用的是这样的短链接模式，先将数据积累在本地，然后每隔一段时间将这些数据递送一次。除此之外，访问网页使用的也是这样的短链接模式，你传送给服务器一个 URL，告诉它你想要这个网址中的内容，服务器找到之后发送给你，但是它并不关心你是否能够收到，它一旦发出就会去处理别的事情，除非你再次发送请求，否则它也不会再来理你。

11.2.2 数据存储

无论是实时的数据采集，还是先本地存储稍后再发送的数据采集，数据最终都要进入仓库之中。我们通常说的数据存储的仓库便是数据库。在大数据时代，传统的数据库已经容纳不下这么海量的数据，因此需要一个更加宽松的大仓库，将数据有序码放，从而更加容易地找到所需要的数据。目前，数据一般被存储在一个叫作 Hadoop 的分布式机器上。

何为分布式？为什么又需要分布式呢？分布二字拆开来看可以这么理解，“分”就是分散，“布”表示及时排布，连起来看，分布式就是将数据分散排布在不同的各个仓库中。这里需要补充的一点是，尽管我们之前使用渣土来比喻数据，但是两者还是有本质区别的。渣土一块就是一块，它如果放在仓库 A 中就不会放在仓库 B 中。但是数据可以被复制，如果它放在仓库 A 中，我们同时也可以复制一份放在仓库 B 中，这样即便 A 中的数据找不到了，也可以在 B 中找到。因此，数据分布存放的好处就是防止数据出问题，即数据的灾备。

是否只需将数据放到相应的仓库中就可以了呢？当然不是，我们还需要知道什么数据放在哪里。为了达到这个目的，我们需要建立一个账本，这个账本也是一个字典，通过查询，知道某些数据存在哪些仓库，某个仓库中又有哪些数据，然后按图索骥再去找到数据。

现在让我们做一个简短的总结，这里涉及了三个对象，我们称之为存储的三兄弟，因为每个存储的过程都需要三者协作完成。首先是数据的递送者，负责把数据运送来；然后是数据的分配者，负责记录来了多少数据，都入了哪些库；最后是数据的存储者，负责实实在在地看好分配给自己的数据。

既然三兄弟各司其职，那么是否意味着三者只需要各守一方即可呢？并不是这样，在存储的过程中，他们三个需要高度配合。典型的配合场景有三种，一曰存储数据，二曰取出数据，三曰数据容错。下面我们就来分别看一看这三种典型场景下存储三兄弟是如何配合的。

存储数据

一个数据到来之后，它需要经历的典型过程可以概括为“切、备、分、传、关”。

切指的是数据到来之后需要进行切分，同类数据应该放在一起，而另外一类数据应该放在别处，除此之外，仓库中摆放数据的货架大小有限，因而数据也需要被切分成可以被放在货架上的大小，这是通过限制数据文件占用磁盘的大小来控制的。备即备份，为了防止数据丢失，于是要将一份数据同时复制成三份，这也是分布式存储的意义所在。分即分配，数据备份好后就可以在账本上记账了，实际等于给数据分配了一个位置。传即传送，分配好数据之后，只需要按照约定位置，一个一个按照顺序送到应该去的仓库即可。关即关闭，最后只需要把仓库的门关上，保证这些数据的安全即可。

取出数据

数据已经在库中了，如果需要找到某个数据并且取走，那么典型的过程则可以概括为“请、查、传”。

所谓请，就是请求账本的保管员，让他帮忙查找所需要的数据；接下来管理员从账本中查找到你所需要的数据信息，这个信息包含数据所在的位置以及大小等，即为查；最后你要拿着这样的信息，到指定的仓库前取出数据，仓库的管理员会将数据传送给你，即为传。

数据容错

无论是存数据还是取数据，都是在没有错误发生的前提下进行的，那么现在问题来了，是否存在妨碍你完成上面两种行为的情况呢？答案是肯定的！这样的情况我们就称之为数据的错误。而针对这样的错误采取的措施就是容许错误存在的保障措施，简称“容错”。

情况一是，某个存储数据的仓库损坏。为了解决这个问题，各个仓库的保管员需要定时向保管账本的人汇报一次，以表明其仓库是完好无损的。一旦发现某个仓库管理员在该汇报的时间还没有汇报，就认为这个仓库已经出现问题了，于是就要采取一些措施进行补救，例如将本来存在这个仓库里的数据转移到其他仓库中。只要知道这个仓库中有哪些数据，就可以从存有这些数据的仓库中再次复制并进行备份了。这种容错方式叫作心跳报送。

情况二是，仓库没有问题，而且取数据的人已经到了相应的仓库门口，但保管员却无法取出数据。为了保证不出现这样的问题，光把需要取数据的信息给保

管员还不行，还需要等到保管员把数据拿到你面前亲手递送给你，才算是真正确认了这个取数据的过程已经完成。

情况三是，保管员取到的数据并不是你要的数据，或者你要的数据可能已经损坏，这种情况下可以通过检查数据中的防伪码（其实是校验码）来确认保管员取到的内容是否是你想要的。

11.2.3 数据计算

数据存储仓库中只能算是压仓底，要想发挥数据的价值还是得多使用数据，正所谓“流水不腐，户枢不蠹”。

数据的使用首先体现在计算上，在大数据时代大家最为津津乐道的就是 Google 出品的 Map-Reduce，而很多关于 Map-Reduce 的介绍都是以 Word Count（统计字符出现次数）的角度来进行说明的。首先来看一下图 11-3 中的一列字母，这一列中出现了多少个字母？每个字母各出现了多少次？

a

b

a

a

b

c

b

思考：如何统计左侧字符出现的次数？

方法1：逐个计数（适用于小数据）

背后机理：从头扫描→添加到计数结果表中→继续扫描→在结果表中查找是否存在？→存在则修改，不存在则添加到结果表中。

方法2：分工合作（适用于大数据）

背后机理：工业社会大分工的必然要求

图 11-3 Hadoop 中的 Word Count 示例

很多人会觉得，这样的问题还不简单吗？甚至有人直接就可以报出答案。但是要知道，这里的列表长度非常之短，如果这个列表再长 100 倍，里面的字符也不会只有 abc，那么就会很复杂。

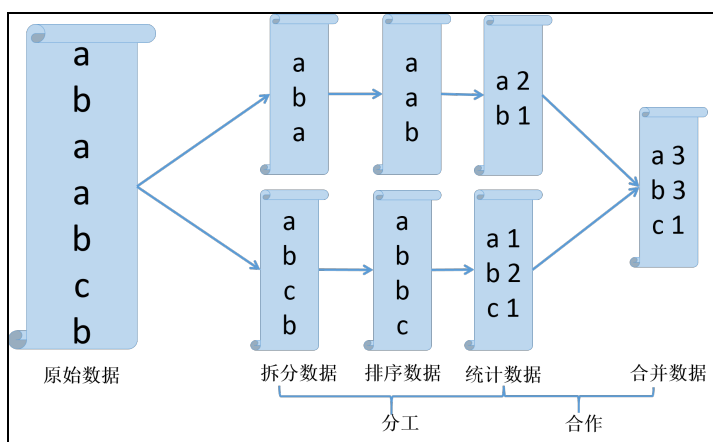
大多数人解决这个问题的一种模式是一种线性的处理模式，即“逐个计数法”。

这样的过程适合小数据，在大数据的情况下，一方面数据大到我们根本没有

办法打开这个列表（计算机的内存无法读取数据），因而也就没有办法计算。另一方面，进行查找更改操作极为复杂，它的复杂程度是与纸带列表中有多少个字母相互关联的。于是乎，人们在思考，有没有办法解决这两个困难呢？

第一个困难比较容易解决，既然很大的数据没有办法一次性打开，那么就使用我们之前介绍的“分而治之”思想进行拆分，把它转化成小数据就可以了。针对第二个困难，由于我们把数据进行了拆分，因而原来整个纸带列表中的字母或者单词数目就被均匀地划分到各个计算机上进行分别统计了，于是进行记录的工作量自然也减少了，可见“分而治之”的策略是如此有效，而这背后的原理则是分工合作。

让我们来看看机器间究竟是如何进行分工合作的。首先，一个原始的纸带列表会被分成两份，我们不妨假设从第三个和第四个字母处进行划分，真实的划分情况远比这复杂得多，但大同小异。这个时候拆分后的数据被分别分配到两台机器上，每台机器首先要对其进行排序（排序的过程前文有述）。这个过程实际上是将计数过程中的复杂性转移到了排序过程中，因为排序过程的算法众多，于是可以在这一步进行单独的优化，使得效率更高。当数据排序完成后，我们只需要进行统计即可。一旦两台机器都完成了各自的统计工作，它们只需要将各自的结果进行合并即可。合并的过程也是及其简单并且省时的。因为统计列表都是经过排序的，因而只需要采用之前介绍的归并排序中的方法就可以完成了。



到现在为止，我想你已经清楚了两台机器是如何分别处理大数据的，不过你或许会问，这与 Map-Reduce 有什么关系呢？其实 Map 就是分工，而 Reduce 便是合作。

11.2.4 数据分析

数据的计算只不过是一个最基本的操作而已，如果把千千万万的计算操作堆积在一起，就会产生一些有意思的操作，我们称之为数据分析。

在我看来，数据分析主要分为两大类：一类是商业智能，例如 OLAP；另一类是数据科学，例如数据挖掘算法等。OLAP 即 Online Analytical Processing 的缩写，是一种比较传统的数据分析技术，其原型是把数据看成一个立方体，或者高维的立方体。在这里我们不妨使用三维立方体进行阐述，这意味着数据具有了三组属性，例如学校、家庭户口、年级，那么此时数据就是学生人数，这个时候的数据立方体可以进行诸如“钻取、上卷、切片、切块、旋转”等操作，简单来说就是查询“各个年级有多少学生”“某个学校的农村户口学生有多少”之类的数据。

而对于数据科学来说，其使用的方法不同于 OLAP 中的简单累加求和，可能还需要用到之前介绍的各种数据挖掘算法。尽管这些算法已经不需要再写复杂的程序，但是数据科学最重要的问题在于如何定义问题以及如何把各个业务领域人员定义的问题翻译成数据挖掘的问题。因而不仅仅需要技术，还需要逻辑处理与创意。

11.3 能聊：跟随销售面向市场找思路

菲利普·科特勒（Philip Kotler）是美国的经济学教授，也是当今营销界的权威人士，至今我的案头还放有科老先生的著作《营销管理》。

提到销售，不知道在你的脑海中其对应的形象是什么，是每天接到的问你是否贷款的电话推销，还是店铺门口扯着嗓门自吹自擂的店员。有些人认为销售就是卖东西，其实这是一种片面的看法。销售人员能够自成一家，并与产品、研发

等并驾齐驱，自然有它的道理。我很乐意与公司的销售人员聊天，每一次和他们一起出差，我总有问不完的问题，学不完的知识。

营销人员是最有激情的，他们每天都要见不同的人，因此需要把自己最好的一面展现出来才能签单有业绩。与之相较，研发人员每日和固定的人打交道，“日久生厌”是常态。

营销人员是眼光敏锐的，他们除了见客户也要见竞争对手。如果把公司内的研发比作造武器的，营销人员便是赤手空拳在前线随机应变的那一个。和他们相比，产品人员走出去较少，在家中“冥思苦想”更多。

下面从营销对象、参与人员、销售渠道三个角度带大家走进营销。

营销对象

我们通常会把业务分成两种类型，一种是面相终端消费者的，称为 to Customer，简称 toC。另一种是面相企业客户的，称为 to Business，简称 toB。在这两者的基础上，从企业客户中又衍生出一类特殊的客户，即政府。由于面相政府的营销与面相企业的营销在收益的即时性以及社会效益角度的考量上略有不同，因此也有人将面向政府的销售单列出来，称为 to Government，简称 toG。

针对这三类销售对象，营销过程有不同也有相同之处。

不同之处在于，toC 的付费者是单个用户，买单人往往就是产品的使用者。要从这样的用户口袋里将钱掏出来何其难也。这个社会中的选择如此多样，用户为什么要选择你而不是别人？因此在面向 toC 用户的时候更多的是要覆盖共性。当共性被极大满足的时候，也就是市场成熟的时候，需要差异化的个性产品来切入或者开拓新兴市场。正如所有人都知道的那样，现在已经不是同类产品之间的竞争了，而是不同行业的产品在不同场景下争夺用户。抢走方便面生意的不是其他方便面厂家，而是在线外卖平台。对于 toB、toG 用户来说，付钱的人和最终使用产品的人往往不是同一个人。例如教育部门为区域内的教师采购教具，付钱人可以是教育部门，可是使用人却是各个学校的老师。在面向这类客户进行销售的时候，需要从其位置进行考量，要明白决策人在担心和顾虑什么。

相同的是，无论是 toC 还是 toB，销售人员最终面对的都是活生生的人。移情和换位思考是最基本的要求，如此才能维护客情关系，便于长久业务拓展。

参与人员

不同的企业生产的内容不同，有的是实打实的产品，如家电行业、汽车行业；有的是看不见摸不着的产品，如软件系统等；还有的是更为虚无的服务业，如咨询行业等。不同的产品在销售上也需要不同的人员参与。

从家电产品到咨询服务，越是从实走向虚，其销售人员和专业技术人员的界限越是模糊。对于家电行业来说，研发归研发，生产归生产，销售归销售，很少看到研发人员上阵做销售。在咨询行业中，除了培训课程开发之外，专业人员也需要直面客户，因而也可以说其是销售人员。

无论是实体制造业，还是咨询服务的第三产业，产品从生产递送到客户手中都有很长的路要走。这里根据路径中参与人员属性的不同有两种主要的销售模式，一种称为直营，另一种则称为代理。直营，顾名思义，就是销售人员隶属于生产产品的企业或厂家，他们直接把产品通过某种途径递送给客户。这样的模式需要企业在各地建立基地，并进行大量的人员培训与管理。对企业来说，这是一个比较辛苦的过程。可一旦建成，其营销能力便不能小觑。代理，则是找到当地的企业，在其帮助下把产品与货物递送到客户或用户手上。这样的代理企业具有很强的区域属性，往往在一个地方人脉深厚，能够帮助产品打开销售天地。

至于选择直营还是代理，这需要取决于企业自己的战略。

销售渠道

无论是直营还是代理，也仅仅是从厂家拿到了货物而已，要与最终消费者接触还需要各种渠道。营销“4P”理论中的 Place 指的就是渠道，意思是用户接触产品的场景。

不同的行业对渠道的划分也略有不同。按照规模来划分，可以分为大与小两种。大型渠道在一些行业中被称为终端，指的是具备一定品牌影响力和综合营销实力的大型购物中心以及卖场，如国际连锁品牌沃尔玛、家乐福、麦德龙等，除此之外还包含一些国内大型购物广场，如银泰、万达、百大等。这些大型终端有一个专业的名称，叫作 KA，即 Key Account（重要账户）。往往这些 KA 会有很强的话语权，因而也比较强势。产品要进 KA，得有过人之处。

另一类小的渠道包括社区或者路边的夫妻店，以及一些零散的企业或政府采

购等。这些渠道为企业贡献的是小流量，或是少量交易，或是低频次交易。

不同的渠道覆盖的范围不同。因而在选择代理的时候，既要考量代理的综合实力，也要考量自身产品特性与代理所拥有渠道的匹配程度。如果一个代理所拥有的是大型企业和政府的渠道，那么进行行业软件售卖时签约这家代理就没错了。但如果代理所拥有的是大型卖场的 KA 渠道，则很难想象把行业软件摆在大商场中售卖是什么情况。

part four

04 第四部分

数据产品经理的自我修养

第 12 章

学习力：借方法论加速

12.1	方法论知多少	266
12.1.1	概念阐述	266
12.1.2	分类总结	267
12.2	学习过程的“满灌”与“脱敏”	269
12.2.1	理解提炼	269
12.2.2	我的方法论	271

方法论这个哲学术语，对于一些人来说是飘在空中的概念，对于另一些人来说却是指导思考的一把利器。用不用方法论，不取决于你所处位置的高低，也不取决于你见识的长短。无论你拥抱还是排斥这个概念，不得不承认，你在平时的工作中一定会潜移默化地运用到它。

使用方法论进行学习，可以起到事半功倍的效果。一方面，方法论为学习的过程提纲挈领。我们遇到的大多数被称为法则、原理或者效应的方法论，往往简短，便于识记，可以作为阅读、思考、分析的框架。另一方面，方法论为学习的内容画龙点睛。许多学习内容如果自己看，往往看不出门道，但是经过别人的总结就觉得甚是有理。因此方法论的思维启迪作用不可小觑。

鉴于此，本章分别从概念阐述、分类总结、理解提炼与我的方法论四个方面聊一聊关于方法论的话题。

12.1 方法论知多少

12.1.1 概念阐述

根据 MBA 智库百科与百度百科的定义，方法论是人们认识世界、改造世界的根本方法。

用途是认识与改造

认识和改造是解决一件事情在时间上的先后步骤，我们对一件事情或一个问题需要先认识后改造。方法论参与的阶段可以是其中某一个，也可以是两个。利用方法论认识世界可以避免冲动，而使用方法论改造世界（即找到解决问题方法）则可以保证逻辑严密、思路清晰。

本质是方法的方法

方法论既然是哲学上的概念，那么其不可避免地“务虚”“理论”等概念划为一类。“方法的方法”听来拗口，但在工作、学习中却时常出现。对于软件开发人员特别是数据库工程师来说，元数据概念深入人心，元数据就被定义为数据的数据；对于国家公务人员来说，他们的工作是治理国家，而治理被定义为管理的

管理。观察方法论、元数据、治理这三个词，不难发现，方法、数据、管理偏向具体操作层面，而方法论、元数据、治理则偏向概念与制度层面，因而方法论可以被理解为一个模板，或者是一个“套路”。

内容是任、阶、技、工

当我们说到方法论的时候，我们在说什么？大致看来，方法论的内容包括了任务、阶段、技巧、工具等。从应用场景来说，又可以分为分析、管理、市场、营销、战略等。这里需要强调的是，无论是内容还是场景，它们并非完全互斥，而是略有交集。

下面列举一些我们平常接触较多的方法论，大家可以趁这个机会检验一下自己平时到底受方法论影响多深。例如 PEST 分析模型、SWOT 分析模型、5W2H 分析模型、逻辑树、MECE 原则、电梯法则、4P 营销理论、用户行为理论、麦肯锡矩阵、波士顿矩阵、十字象限分析法、波特五力分析模型、波特价值链分析模型、SCP 分析模型、STP 理论、GTD、6 点优先工作制、帕累托原则、莫法特休息法、六顶思考帽、生命周期理论、金字塔原则、SMART 原则、kano 模型、成熟度模型等。由此可见，方法论大多以模型、理论、原则、制度等词为中心词，这也恰好印证了我们之前讨论出的“方法论的本质是模板与套路”这一论断。

12.1.2 分类总结

以上所有的方法论都是在自我或企业分析与管理、市场营销或战略制定中分别提出的，因而有的虽然名字不同，但是本质是一样的。接下来，我们将对这些方法论进行归纳总结并分别阐述。

按应用场景进行分类

按照应用场景对这些方法论进行分类，可以分成分析管理类、市场营销类、企业和个人战略指定类，其中包含的方法论分别如下。

- 分析管理：5W2H 模型、逻辑树、MECE 原则、电梯法则、十字象限分析法、GTD、6 点优先工作制、莫法特休息法、六顶思考帽、金字塔原则、SMART 原则。
- 市场营销：4P 营销理论、用户行为理论、波士顿矩阵、生命周期理论、kano

模型、成熟度模型。

- 战略制定：PEST 分析模型、SWOT 模型、麦肯锡矩阵、帕累托原则、SCP 分析模型、波特五力模型、波特价值链分析模型。

按涉及的内容进行分类

当按照这些方法论的内容进行分类时，可以把它们分成解决特定任务类、涉及时间阶段类、处理事情通用技巧类以及常用工具类。

- 任务：金字塔原则、4P 营销理论、用户行为理论、波士顿矩阵、kano 模型、麦肯锡矩阵、SCP 分析模型、STP 分析模型、波特五力模型、波特价值链分析模型。
- 阶段：生命周期理论、成熟度模型。
- 技巧：5W2H 模型、MECE 原则、电梯法则、十字象限分析法、6 点优先工作制、莫法特休息法、六顶思考帽、SMART 原则、PEST 分析模型、SWOT 模型、帕累托原则。
- 工具：逻辑树、GTD。

按目的与形式进行分类

上述这些常用的方法论很多是可以合并的，我们根据其目的或者形式进行合并，可以更加深入理解这些方法论之间的联系。

- 与时间管理目的相关：GTD、6 点优先工作制、莫法特休息法、帕累托原则、电梯法则。
- 与矩阵分析形式相关：波士顿矩阵、麦肯锡矩阵、十字象限分析法。
- 与战略决策目的相关：PEST 分析模型、SWOT 分析模型、SCP 分析模型、STP 分析模型、波特五力模型、波特价值链分析模型。
- 与逻辑组织目的相关：5W2H 模型、逻辑树、MECE 原则、SMART 原则、六顶思考帽、金字塔原则。
- 与抽象建模目的相关：4P 营销理论、用户行为理论、生命周期理论、kano 模型、成熟度模型。

更多关于每一条方法论的详细内容可以在互联网上轻松找到，这里就不再赘

述了。

12.2 学习过程的“满灌”与“脱敏”

学习是对学习内容感兴趣并逐步掌握的过程。一下子介绍了这么多方法论，要想在工作中灵活自如地运用实在不易。通常来说，有两种主流的学习方法可以解决这种困惑。第一种称为“满灌”（Flooding），即通过短期内的强化练习让学习者增加与学习内容交互的频率。第二种称为“脱敏”（Desensitization），即通过在长期的工作过程中间歇式地与学习内容进行交互，从而产生潜移默化的影响。

上述两种方法源于心理学家华生（John H. Watson）创立的行为学派，而理论基础则源于巴浦洛夫（Ivan Petrovich Pavlov）的条件反射实验、桑戴克（Thorndike）和斯金纳（Skinner）的操作性实验以及班杜拉（Bandura）的社会学习理论。在行为主义理论之后，又相继出现了认知主义与建构主义，后两者的理论不再仅仅关注外界条件和行为结果之间的联系，而是将重点转移到了人的认知与反思过程上，以及与外部环境的相互影响上。

我们所熟知的、经久不衰的方法论大多来源于他人，会转述、会用是最基本的要求。如果在此基础上，能够将这些方法论拆解，并重新对其形成认知，便可以在以后的工作中总结出属于自己的方法论。

12.2.1 理解提炼

我们可以对各个类别的方法论做一个简单的提炼，看看这些方法论都有怎样的特点。

时间管理类方法论，要求简单实用。对于矩阵分析类，需要直观易懂。对于抽象建模类，需要有一条主线。对于逻辑组织类，名字一定要严谨清晰。对于战略决策类，往往是为了分析市场、行业等宏观环境，那么就需要全面且具体。

用颜色、数字或字母来命名方法论，是为了朗朗上口。用逻辑来组织方法论则是为了铭心刻骨。方法论的内部往往是由多个方面组成的，这些方面之间有着很强的逻辑关系。这些逻辑关系可以是并列、承接、递进、转折、因果、条件、

假设、诠释等。使用这些逻辑关系组织起来的方法论在传播上更加深入人心，难以忘记。有一项研究表明，人能够一次性记住的不相关事物不超过 7 个，通过逻辑关系把这些事物组织起来，则能够大大提高记忆内容的体量。

这样看来，方法论并不神秘，甚至造一个方法论都有了一定的思路。那么什么才是评价方法论的标准呢？换句话说，什么方法论才是好方法论？我的理解是，“好用”是检验好方法论的唯一标准。无论是完成一次汇报，还是向同事解释一个业务，或是组织一个项目，都可以通过在这些任务中使用方法论并评估别人的反馈来检验某个方法论是否好用，是否需要根据实际情况加以定制和改造。

初入职场的同学或者刚刚接触方法论的同学，可能会觉得别扭。这样的别扭来自两个方面，一方面是因为陌生知识领域的门槛较高，往往有望而生畏之感，而且方法论从概念上讲比较虚，难以在工作中拿捏，很难灵活运用。另一方面，方法论就是套路，现在不是流行“少一点套路，多一点真诚”的说法吗，从内心对套路的排斥也可能是初入职场的小伙伴们的情怀。

我想说，方法论的套路与心中的情怀，并不矛盾。我们常常在网络博文中看到关于“道”与“术”的讨论。讨论中提到，“术”是低级的方法，“道”是高级的方法。方法论可能就介于两者之间，然而情怀则是高于“道”与“术”的另一境界，即在混沌中解决问题于无形，不需要套路的红绿蓝，也不需要步骤的一二三，就能够把问题解决。

“混沌”看似玄妙，其实不然，当然也不是可以一步登天的，需要在运用方法论的日积月累中慢慢固化，乃至内化，才能达到这样的境界。任正非曾在描述华为流程管理的进化历程时提到“先僵化、后优化、再固化”。我觉得可能还得再增添一个“内化”，即消化成自己的知识并表达出来。这也是我们学习一个新知识，步入一项新领域，应该有的步骤。一开始，我们只能蹒跚学步，后来我们开始优化与改进，接着当我们把优化的步骤、流程或者套路在实践中不停检验，发现非常好用之后，就会希望这个套路能够成为一个标准化的流程，于是就固化了下来。进而，由于你是固化了这个流程的人，心中难免会有一些“矫情”的心路历程、感悟，这其实就是内化的过程了，说明你已经能够上一个台阶去看问题了。事情是一点一点做的，方法论也是一点一点累积的，慢慢来，不用着急，这几步缺一不可。

12.2.2 我的方法论

在这里，我主要想谈一谈我个人浅薄的认识，关于如何学习一个新领域的知识。根据我的一些经历，无论是学术上，还是工业界中，一般来说，书籍、报刊、博文等的阅读量达到一定的量，这个领域的知识差不多就已经了解全了，这个量一般不会太大，在100篇左右。不过，这些文章需要短时间内集中阅读，并且要保证内容中的重复和相似部分较少。当继续阅读的时候，差不多会有这样一种感觉：好像这篇文章讲的东西我见过。这种感觉越强烈，越能说明该领域的知识已经开始被穷举了。

紧接着，我会把文章中的要点提取出来，进行分类和总结。总结的过程大概会以两三篇写得较好的综述为大框架，然后以其他文献作为补充，向大框架中填充。等这个工作完成，一个属于自己的对于该领域的总结就完成了，并且我相信这会加深自己对该领域知识的认识。

然后，进行发散思考，联想其他行业以及领域的相似性。我们都明白一个道理，所有的领域都是相通的，我们时常会感到某个领域的情况或者定律好像在其他行业也有类似的对应。总结得越多，越能够发现，其实各个领域知识之间有很强的关联性。这是发人深省的好时机，多做发散思考，多思考为什么会有这样的相似，可能就会在一些无稽的想象中萌发出一个概括性的普世解法。

最后，找出关联并提出主张。在这个阶段，你已经可以融汇贯通多项内容，于是可以根据上述总结的原则，给出一个你所想象出来的名字，冠名属于你自己的方法论。把领域和领域之间的交集找出来，做好联系，在脑中构建一张图，以后需要某个领域知识的时候，说不定就可以借用一下其他领域的方法。

很多知识与内容都是重复的，也是枯燥的，但却是有用的。如同华为强调要将制度先僵化，再优化，后固化，知识也是如此，刚开始进入一个领域的时候，我们不妨怀着崇敬的心情去吸收每一篇文章给我们带来的知识享受。慢慢地我们要学会挑选并且针对文章中的观点做一些思考以达到去伪存真的效果，这样才能使知识固化在我们的骨髓中。知识不仅仅要固化下来，还应该内化，而内化的本质就是可以举一反三，就是可以对外输出。这也是我写这本书的初衷所在。

本章在写作的时候也遵循了这样的主线，即罗列、分类、总结、发散、联系这几个步骤。大家可以根据自己的学习方法，总结出属于自己的方法论。

第 13 章

表达力：用逻辑学帮衬

13.1	写得一手好文案	274
13.1.1	为公务员考试正名	274
13.1.2	写作实战简明教程	275
13.2	讲故事给同事听	278

尽管我们曾提到过“优秀的数据分析会说话”，但作为数据产品经理，还是会在工作中经常遇到和人打交道的情况。“酒香还怕巷子深”，不会表达严重影响了整体工作成果的呈现，会造成“茶壶里煮饺子，有货倒不出”的窘境。因此，学会表达十分重要。

表达主要分为两种，一种是文字表达，一种是言语表达。无论是其中的哪一种，都需要以逻辑作为第一前提。下面就让我们来学习这两种表达力。

13.1 写得一手好文案

对于大多数人来说，这是一个老生常谈的话题。我们从小学开始就被要求写各种文章，议论文、记叙文、抒情状物文，应有尽有。但是可能很多人直到看到这里也不敢承认自己的写作水平，特别是在平时工作中需要写文档，写方案时，常常会感到力不从心。

13.1.1 为公务员考试正名

公务员大家并不陌生，但是公务员考试却鲜有人问津。公务员考试不同于其他技能类或者证书类考试，似乎只有致力于从事公务员事业的人才会参加。

一些在企业工作的同学往往对公务员考试不屑一顾，他们或是认为这样的考试太过程式化，或是认为未来从事这样的工作缺乏激情。然而事实真的是这样吗？

一个偶然的的机会，当时即将毕业的我参与了上海市公务员局组织的一场面向全国若干所高校的公务员招考，我顺利通过笔试，进入面试，并最终拿到 offer。为了这次考试，我准备了一个月。通过这一个月的准备，我接触了公务员考试的实质内容。首先，认为考试程式化本身就是一个曲解。听过很多人说“公务员考试是八股”，而事实上有这样认识的人往往很难拿到高分。公务员考试的内容并非官方原创，如果你在备考前也接触过 GMAT 考试（美国经管类研究生入学考试），便会发现两者的考题类型相似度惊人，甚至有些题目直接译自 GMAT 真题。其次，公务员考试的两个部分“行测”与“申论”都是对逻辑的考察。从选择恰当的词语到语句通顺排序，从根据一段话推测结论到结合场景书写报告，无一不是在考

察考生的逻辑、条例与应变能力。这些都是在实际工作中真实存在的场景，由此可见，公务员考试是一个实用的、务实的考试。我必须得为其正名。

正因为公务员考试中有“申论”这样一个测试环节，所以当我们谈论文案写作的时候不可能不谈它。“申论”源自于孔子的“申而论之”，其中“申”表示申述，即提出观点；而“论”则是议论，论证观点。因此申论的文案写作更倾向于议论文，需要更多的逻辑与结构化表达。

在结构化表达方面，我最推崇两类书籍。一类是由人民出版社出版的国家领导人重要讲话文章选编类书籍，如《习近平谈治国理政》等。这类书籍的编纂十分用心，仅从目录便可以清晰了解书本结构，且逻辑上相近的章节紧邻在一起，便于查阅。进入单篇文章中，文章的结构清楚，行文言语精炼，概括总结到位，语句排比对仗工整，堪称范文。另一类是由国际知名咨询公司麦肯锡出品的诸多教授结构化思维的书籍，如我们提到多次的《金字塔原理》《麦肯锡教我的写作武器》《麦肯锡教我的思考武器》等。这些书籍中系统介绍了结构化思维的技巧，难怪麦肯锡咨询的报告能够被广泛认可，其核心秘密就在于逻辑严明。

13.1.2 写作实战简明教程

我的职责之一是编写方案以及申报课题或项目，在此过程中总结和积累了一套属于我自己的写作方法，且让我与你分享。

了解几个思考模式

在进入正题之前，让我们先了解和回顾一下思考的若干种模式，分别是演绎、归纳与罗列。

演绎主要是通过一个基础的论点展开推论，通过举例的方式进行论证。例如“水果可以补充维生素”这个论点，使用演绎法进行扩展就可以变成“苹果可以补充维生素”“橙子可以补充维生素”。由此看来，演绎是一种由一般到特殊的方式。

对于归纳来说，则是从多个实例中提取共性找到论点。例如我们找到了家庭条件不好的张三，发现他很小的时候就懂得帮助父母做家务。又找到了同样家境窘迫的李四，发现他学习之余还勤工俭学补贴家用。于是我们就可以得出“穷人

的孩子早当家”这样的结论。因此，归纳法是一种从特殊到一般的方式。演绎和归纳是一对彼此相反的思考过程，前者是自上而下，后者则是自下而上。

除此之外我们还可以通过平行的构建来进行思考，这就是我们常说的罗列，罗列的前提是满足 MECE 原则。例如我们为了设计一套教育软件，就需要根据用户类型进行细分，于是可以列举用户类型为学生、家长、教师、校长、教研员、教育管理者。他们中两两之间并不交叠，所以满足了 ME (Mutually Exclusive) 原则。他们之间的交集又覆盖了几几乎所有的用户类型，因此满足了 CE (Collectively Exhaustive) 原则。

确定一个写作结构

接下来我们就要看看如何进行写作了。

我们在写文章的时候经常按照以下思路来进行，先是介绍背景，然后引出冲突并提出疑问，最后解决。可是为什么要这样写呢？换言之，这样写的好处是什么呢？不知道你注意到没有，通过先背景，后冲突或挑战的模式，可以制造一个高潮。鲁迅先生在《再论雷峰塔的倒掉》中指出：“悲剧将人生的有价值的东西毁灭给人看，喜剧将那无价值的撕破给人看。”因此通过制造冲突可以吸引眼球，从而获得关注。

在这几个环节中，背景可以是一个大环境的介绍，大趋势的介绍，或者是通俗的小故事的引入，甚至刚刚发生的时下热点或身边事也可以作为导入。这部分内容相对来说和大众的认知吻合，因而容易获得认同感。接下来为了抓住用户的好奇心，需要质疑一些潜在的谬误或是指出一些明显的挑战，使得用户心头一紧，心里犯起嘀咕。这个时候用户的情绪被拉到一个高值，随后而来的问题定义则使用户的心舒缓一些，告诉他只能解决某一个或几个问题。最后通过方法与过程使得用户的情绪彻底得以疏导和解脱。这种制造高潮再释放的过程很容易获得关注与认同。下面我们来看看，在这几个过程中，我们分别应该关注什么。

首先在背景部分，按照上面所说，介绍的背景应该是让人们一下就能认同和了解的，这非常重要。所以很多人都是从故事开始，从一个大家熟悉的人物故事或者日常生活常识开始，然后再抛出自己的观点，这一招在网红以及自媒体人中比较常见。用户看完这个背景之后可能会产生一阵疑惑，这个时候说明他已经上

钩了。

接着，你需要能够指出一些和他的常识认知有偏差的结论或者现象，这些结论或者现象被称为冲突或者质疑。通常我们说一本小说好是因为其中的人物冲突多，情节自然跌宕起伏。而介绍冲突的三种模式分别是“人无我有”“人有我优”“相较难下”。人无我有很好理解，给人出乎意料之感。人有我优则是稍次一些的方式：以前也有人做过，也有人提出过，但是我有不一样的解决方法或想法，要不要听一听，这或许也可以吸引一些用户的眼球。相较难下是把一个两难的问题抛给用户，让用户纠结起来，当他纠结的时候，他就会思考你会给他怎么样的答案，于是也可以顺藤摸瓜引出下面的解决方法。

接着就需要释放用户的情绪了。我们可以用三种类型来解决用户心中的疑惑。对于人无我有的情况，不妨直抒胸臆，直接说出你要做什么，并给出具体的指示。对于人有我优的情况，给出你的解决步骤，这个步骤中可能暗含时序感，其目的是说清楚你的解决方案与前人的路径有何不同。对于两难甚至多难的情况，需要给出你的选择，并且在接下来的步骤中给出详细的解释。

最后就是给出整个解决过程了，如果说上一步是给出了论点，那么这一步就是要给出论据了。总结要遵循“总分结构”来进行。在总结句中要遵循 SMART 原则，切勿使用“我将介绍以下三点”作为开头，因为这将没有任何的信息量。在分别阐述的过程中，需要从结构、用词与内涵三个方面来斟酌。对于结构来说，可以使用示意图、编号、加粗、缩进与强调（如字体和背景色）来突出文章结构。对于用词来说，最好保证句式统一，如两个并列句都使用名词性短语做主语或者使用动宾短语。除此之外，引入新词也需要注意，要尽量避免为了解释一个概念而使用若干个新概念，这种嵌套式的解释过程体验非常不好。在内涵上，需要简练不要啰嗦，为了让用户能够听明白，可以适当“模糊”处理某些论据。虽然这样的模糊处理在科学上站不住脚，但是却有利于听众从一个模糊的边缘以及其本来就具有的认知层面上切入，待其进入这个领域后，再加以细化不迟。切勿为了精准而解释一个冗长概念。当然在内容介绍的时候，需要有一个完整的主线。

掌握若干扩容技巧

很多人觉得工作中的文案既不是抒情文，也不是状物文，难以发挥，因此没有什么可写的。我曾经在新东方短暂地从事过一段时间的托福写作课教学工作。

在上岗前的磨课环节，需要每位应聘教师准备一个知识点的讲解。我当时在一个好友的帮助下准备了“英语长句的写作”。

英语的简单句很好写，基本上 5~6 个单词就可以描述清楚一件事。但是这样的句式既不能反映考生“渊博”的学识，也看不出考生“智慧”的头脑。因此简单句需要被扩展成一个约 150 字左右的长句。这个长度基本上和一个小段落的长度相同，比起原来表达本意的句子，长度长了整整 30 倍。英语语句中的扩容很简单，先通过时间状语从句和目的状语从句，把简单句发生的背景交代清楚，然后为简单句中的主语和宾语分别加上修饰语或定语从句用以解释说明，最后通过表示让步或承接的词语，连接一个并列句用以抒发情感。简单三步，一个英文简单句就变成了复杂句。

在中文写作中列好提纲犹如打好了草稿，是完成写作的前提。在具体描写每个点的时候，若能掌握方法则可事半功倍。我自己使用的方法主要也分三步，即在介绍某个观点的时候，分别写清楚是什么(What)、为什么(Why)、怎么做(How)。是什么就是给观点下定义或者阐述概念，讲清楚这个概念的发端与周边联系。为什么则是要讲述这个概念和观点的意义，以及这么做的原因，这个意义又可以从多个维度（如经济、社会等）来展开。怎么做，就是给出详细的行动计划和具体步骤，在介绍具体步骤的时候可以按照总分的结构来进行。

在上面的行文过程中，无论是内容介绍时候的完整主线，还是跌宕起伏的用户情绪，可能都会在一篇文章中多次出现。有的时候上述技巧是嵌套出现的，但任何时候都别忘了写一篇文章的本质是什么，究竟是通过故事普及知识，还是凭借常识引出新思路。无论是哪一种，无论在过程中使用什么技巧，都应该以终为始。在此基础上，若还能结合当下的热点那便再好不过了。

13.2 讲故事给同事听

法国大革命后的 19 世纪，格林兄弟出版的《格林童话》，陪伴的远不止数以亿计的儿童，也陪伴当时的德意志人民度过了黑暗的长夜，等来了民族崛起的朝阳。这一切，都是故事的功劳。

在现代，做汇报需要讲故事，做产品需要讲故事，做销售需要讲故事，做融

资更需要讲故事，我们无时无刻不在接触各式各样的故事。人们如此痴迷故事，主要原因有两个。一方面是故事能够带给人经验，让人们把繁重的劳动经验总结出来加以传播。另一方面，故事给人带来希冀。无论是上天堂的故事，还是诺亚方舟的故事，都是给人们描绘了苦难的解决方案。正是这些方案使得人们能够获得精神的动力，持续前进。

故事如此重要，以至于每个人都得学会讲故事。在物质极度丰饶的今天，既有国内外的书籍介绍故事思维，也有企业专门针对演讲培训故事思维。下面是我通过观察、阅读和实践所做的阶段性总结。

故事的分类

从文学的角度来说，故事主要分成虚构类（fiction）和非虚构类（non-fiction）。

虚构类的故事创作给我们的启示往往是技巧性的。我们可以借助虚构类的技巧，将平常生活中的多个发生在自己身边的故事加以撮合，并最终通过一个虚构的故事呈现给用户，故事中的人物可以不真实存在，也可以将多个故事中的事件与人物属性加在一个人身上。正如销售人员往往会在对外宣讲的过程中，将自己曾经在不同客户上实践过的不同组件模块组合成一个整体的解决方案，并声称曾在某个虚构客户身上实践过。这也是借助了虚构类故事的描述方式。虚构不是说谎，是源于真实生活，又高于真实生活的艺术创作。

另一类非虚构类则实事求是地描述了历史时间中真实发生的事情。

非虚构类的现实意义是明显的。在陈述方案计划进度的时候，需要按时间顺序讲述，让人们有按图索骥之感。如若想把项目或者方案抽象拔高，则需要以主题为主线，通过总结获得若干主题，并按照主题分别描述。由于非虚构类的故事相较虚构类的故事而言更加真实，往往是讲述者亲历的一手资料，因此在讲述的过程中，陈述者可以更自信地描述更多的细节，也能够配合故事有更多情感和表情的融入，这对于故事中信息的传递至关重要。也难怪，在企业中，同一个 PPT 让不同的人讲，效果完全不同。

论述的结构

讲故事就像是打比方，通过通俗易懂的方式让人们接收观点。创造一个故事与写作在结构创作方面有很多相似之处。除了归纳法、演绎法、并列法外，还有

一种能够打动人心的方式，称为“黄金圈法则”。

前文在介绍扩容的时候，提到过是什么（What）、为什么（Why），以及怎么做（How）的三步方法。在正式性的场景下，这样的平铺直叙没有问题。但场景若变化成说故事，这样的方法就略显乏味。“黄金圈法则”则是稍微调整了三者的顺序。简单来说，黄金圈是一个三层相互嵌套的结构。最里层是 Why，即阐述为什么要做某事，是原因，是价值，更是意义。第二层是 How，这一层阐述了实现 Why 所需要的路径和方法。路径的规划既需要分解的能力，也需要调研和全局把控的能力。最外一层是 What，即从意义出发，有了实现路径之后，具体呈现的结果是什么。利用这种方法，可以引人入胜。

黄金圈法则有很多的具体应用。企业树立愿景和价值观，等价于 Why，实现路径 How 则是战略规划，而一个个具体的产品则是呈现形式 What。在销售场景下，如果以成本为中心来定价并进行产品售卖，本质上还是围绕具体的产品（What），介绍产品的价值（Why）以及生产制造过程或工艺（How）。这个过程是为定价找理由、做解释。而使用黄金圈法则来介绍产品，则是将具体产品放在最后介绍，以制造产品的初心和意义开场，通过意义和价值观得到用户的认可，转而再介绍产品，等同于把成本定价模型转换成了用户的心理定价模型，是一个质的飞跃。

故事的模式

故事的模式和故事的结构不同，结构是对所讲述故事的整体逻辑的描述，而模式则是对具体故事情节的概括。“纸上得来终觉浅，绝知此事要躬行”，我自己总结了用得比较顺手的三个讲故事模式。在一个故事中，三个模式中的一个或者多个都可以出现。

第一个模式是“平凡中的不再平凡”。

我曾在公司为校招新员工安排的分享会上讲过一个故事，这个故事用来说明一个核心的概念“信任”。如果仅仅是进行道理的陈述，如“我们一定要相信自己，相信队友，相信领导”，很难起到效果。于是我当时讲述了一个自己在卧龙岗跳伞的故事。我从坐飞机开始讲起，大家都觉得很稀松平常，因为每个人多多少少都有乘坐飞机的经历。当我指出这个飞机是一个迷你小飞机，只能乘坐不到 20 人的

时候，大家觉得有点新鲜。当我介绍到这个飞机飞到半空中，飞机舱门被拉开之后，大家以为我遇到了特殊情况，纷纷表示惊悚，因为这样的经历并非每个人都有过。最后我在屏幕上投影出一张教练带着我跳出飞机外的照片时，大家才恍然大悟。在这里，我讲述的是我把生命托付给一个不认识的陌生人的“信任”故事。在这个故事中，平凡的是乘坐飞机，而不平凡的是跳出机舱。一步步地诱人深入，最后以小见大，说明了一个浅显的道理。

第二个模式是“辉煌下的痛苦遭遇”。

冰心在《繁星·春水》中有这样的一首小诗：“成功的花，人们只惊羡她现实的明艳！然而当初她的芽儿，浸透了奋斗的泪泉，撒遍了牺牲的血雨。”诗中人们的状态是大多数人所经历的，我们平时读成功的故事多，但奋斗和失败的故事少。人们之所以会对成功痴迷，是因为只有成功的故事才会被报道出来，而失败的故事很少有人问津。人们的好奇心和“八卦”心理总是会驱使自己想要知道真相，或者现象背后的故事。在故事的分享中，把成功和辉煌背后的那些痛苦经历与遭遇讲给别人听，会起到很好的效果。心理学中有一个理论强调，一个公众人物不用害怕犯错和暴露事实。轻微的错误和事实不仅不会让其自毁形象，反而会让其显得愈加真实，愈发可爱，愈受欢迎。

第三个模式是“努力后的贵人相助”。

天道酬勤是中国古代推崇备至的故事，用来说明努力就有收获，在这个故事中天道就是贵人。我们现在常说正能量，喜欢能够激励大家的人，没有人希望身边的人是泄气的皮球。讲故事最难得的就是给人盼头，一个曲高和寡的道理不如一个下里巴人的故事。未来的期许也许对听众来说遥不可及，但是当下的努力确实是可以看得见的。而且有案例证明努力就有回报，谁会不愿意尝试呢？在公司的汇报过程中，或是激励团队时，可以借助这类故事模式。通过陈述曾经成功的项目与当下状态的相似，从而让听众产生信心，相信蓝图，进而起到打气与激励的作用。

上述三个模式都用到了对比的方法。平凡与不平凡，辉煌与痛苦，努力与成功，所有的故事似乎又都是同一个模式，即产生反差。只有反差才会引人思索，也只有不同才会让人印象深刻。

故事的引申

当我们的故事不再是通过文字记录，而是需要鲜活的人来演绎时，个人的魅力就显得尤其重要。根据“梅拉比安法则”，肢体语言、语音语调以及说话内容在沟通过程中的信息传递占比分别是 55%、38%以及 7%。原来内容是如此的廉价，而肢体语言和语音语调才是重头戏。

因此，为了不给自己精心准备的故事扣分，注意着装、在乎谈吐很有必要。演讲者站在台上，一言不发，别人便可以从你的身上读到他们想读的信息了。如果你想传递的是专业性，就该像个精英。如果你想传递的是自由与随性，那就衣着宽松。选择与内容匹配的言谈举止便是准备故事时需要额外注意的内容。

第 14 章

领导力：以经济学诠释

- 14.1 事情背后的选择..... 285
 - 14.1.1 选择价值链上游：剪刀差效应..... 285
 - 14.1.2 学会审时度势：美林时钟..... 286
 - 14.1.3 谨慎选择别人的经验：推绳子效应..... 286
 - 14.1.4 平衡是一个难题：萨伊定律与凯恩斯法则..... 287
- 14.2 人员之间的协同..... 288
 - 14.2.1 你闪开，让我来：绝对优势与相对优势..... 288
 - 14.2.2 无条件开放：零和博弈与合作共赢..... 289
 - 14.2.3 教会团队成员什么是沉没成本..... 290

当尝试探究领导力背后的起因时，我发现，原来领导力和经济学原理的耦合早已植根于历史的土壤中，并且不断发展与壮大。纵观历史上的几次科技革命，从蒸汽时代到电气时代再到信息时代，组织发生了许多变迁，而这样的组织变迁正犹如我们职业的发展历程一样，从个体变为一个小组，然后经历大组织，最后形成扁平化的自组织。

在蒸汽时代，农业文明首次向工业文明过渡，出现了从粗放型经济到集约型经济的转变，而这样的转变背后则是经济的规模效应在起作用。工业文明的分工开始发达，一群人聚集在一起形成了小组。

随着科技的不断发展，人类的交流与贸易更加的发达，贸易必然带来利益，利益驱使社会进行竞争，而竞争最终的结果必然是生物学定律中的优胜劣汰。于是经过了一段时间的躁动期，人们围绕在利益最大化的考量之下，将原来落后的产能淘汰，并进而整合为大型的公司或组织。

随着信息学带来的科技进展，任何信息都可以通过社交网络或者是互联网迅速传遍世界各个角落。每个人平等获得信息，每个人都可以独立进行决策。

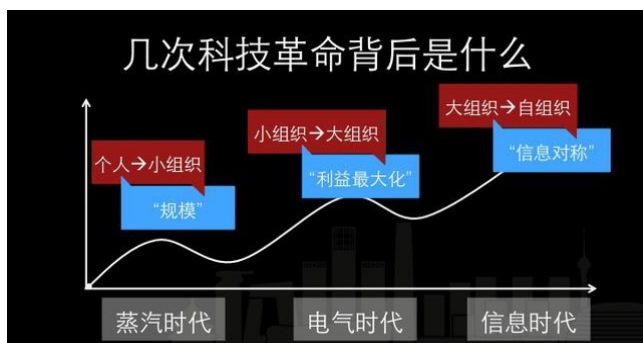


图 14-1 几次科技革命中团队的迭代与升级

可以毫不夸张地说，社会生活背后有一只无形的经济学之手在操纵着我们。作为数据产品经理上升道路上一定会经历的从技术、业务到管理的转型，领导力建设刻不容缓。

帮助团队选择方向、制定流程、评估结果，我们可以把工作做得更好。团队内部的沟通和外部的沟通齐头并进，我们可以把工作环境营造得更舒适。面对突发状况的应急策略则会让我们更从容。这其中每一处所需要的领导力，都可以从

经济学原理中获得。

14.1 事情背后的选择

“领”有率领的意思，也许不一定要求你身先士卒，但思想必须冲在团队所有人的最前面。从方向的选择到时机的选择，从经验的选择到价值判断标准的选择。每时每刻，总有新的事情在等待着你的抉择。

14.1.1 选择价值链上游：剪刀差效应

经济学中有一个名词“剪刀差”。上世纪 20 年代的苏联在走上和平建设道路后，大力发展工业。为了给工业发展聚集资金，于是人为地压低农产品收购的价格，从而使得整个交易过程中工业产品与农业产品的价格与其价值不匹配。当物价上涨的时候，工业品价格上涨幅度大于农产品价格上涨幅度；当物价下跌的时候，工业品价格下跌幅度也小于农产品价格下跌幅度。当我们将一段时间内的工业品价格和农业品价格画在图上的时候，两个价格趋势曲线就会呈现一个张开的剪刀形状，因此工农业产品的价格差被称作“剪刀差”。

这种价值不对等情况在工作中也屡见不鲜。

作为一个 Leader，首先需要考虑的就是团队的工作内容。在数据科学领域，团队中一定少不了数据分析师。数据分析师对于处理报表类的工作自然是应对有余，但是如果更上一个层次就难上加难了。究其原因，数据分析师的技能瓶颈在于其做不了更加复杂的挖掘工作。团队每个成员的瓶颈束缚了 Leader 在规划业务时的手脚。成员能力得不到突破，就只能做些低端业务，越是做低端业务，成员就越难有能力的突破。这仿佛进入了一个恶性循环。类比“剪刀差”，低端业务仿佛是农产品。

为了谋求业务上的精进，作为 Leader 的你需要能够为团队指明一个清晰的方向，说到底还是要选择一个更加核心的业务或是一些创新业务。没有壁垒和护城河的工作总有一天会被淘汰或替代，于是只能不停地向上求解。

在数据科学团队中，可以通过内部集训培养一批数据分析师，使其成为数据挖掘工程师，也可以通过外部招聘渠道引进一些具有算法实践经验的人才。上述

方式都能够使我们的业务转型。“落后就要挨打”是颠扑不破的真理，以前为了避免“剪刀差”而选择价值链的上游，现在在工作中为了给团队每个成员带来高净值增长，我们要培养每个人的能力。

14.1.2 学会审时度势：美林时钟

知道了选择什么样的技术与业务还不够，还需要了解该在什么时候选择，即选择的时机。选择时机是经济界最为津津乐道的话题，他们把找准投资的机会称为“投机”。

美国知名投行美林证券（Merrill Lynch）发明了投资领域最为知名的资产配置理论“美林时钟”（The Investment Clock），用以告诉人们在经济处于什么状况时进行什么类型的资产配置。

无论是长期与短期的平衡，还是合作与竞争的选择，都离不开审时度势。经济学中的美林时钟告诉我们在什么时间该做什么样的资产配置，实际上是在市场经济环境中树立了风向标，教导我们看着风向标行事。

无论是企业内部的团队，还是复杂市场中的创业公司，都需要把握市场的脉搏，顺应行业的风口，选择合适的赛道。在公司施行大数据和人工智能战略的时候，要能够让团队把握住公司的上升势头。在内部项目立项或团队间赛马时，可以针对性地与主题贴合。在外部创业时，也可以针对性地设置项目，以获得融资。如果时代的热点未能出现在工作目标中，很难说这是一种明智的决定。因此作为Leader不仅要具有勇气，还要有技巧，顺势而为。

14.1.3 谨慎选择别人的经验：推绳子效应

绳子是一个很有趣的物体，当你拉它的时候，它会向你的一侧移动，但当你推它时，它却不会向前挪动半步。从物理学上很容易找到解释，因为绳子是柔性的，因此推力难以传导。

推绳子效应在工作和生活中有很多应用。大致是这样的模式：有正反两种情况，为了应对这两种情况，也有正反两种策略。当在正面情况下使用正面策略时，

问题被推向反面，且收到立竿见影的效果；而在反面应用反面策略时会将其推向正面，但却未能收到应有的成效。找找身边是否存在这样的案例呢？

团队的管理就是这样的一个例子。我们常说“一管就死，一放就乱”，而实际情况是“一管就死，一放不敢乱”。

团队的工作模式有两种恰恰相反的极端情况：一种是极度守规矩，按流程办事，凡事需要请示、汇报、批复、执行；另一种则是极度灵活，遇事随机应变，但也经常打擦边球。小团队行事作风偏向于后者，但随着团队的扩张，人数的增多，为了将每个人管理起来，降低工作中的风险，就需要把人放在流程上，更多依靠流程。这是现代管理制度最伟大的发明。慢慢地，团队开始向第一种工作模式转变。人数的增多使得以前能够朝夕相处的人每天不一定能说上一句话，本能通过当面沟通解决的事情却要通过邮件你来我往，降低了效率。但当这样冗余的流程和复杂的管理变成了负担和成本时，团队的管理者便又希望团队能够转向第二种工作模式，开始二次创业，激发大家的激情。“然而大家已经习惯了邮件确认工作，习惯了明确责任的边界，再想把思想换个方式就不像开始那么容易了。在软件开发过程中，瀑布流开发和敏捷开发也是类似的一对概念。

好比推绳子效应，从一种状态转变到另外一种状态很容易。这种容易有可能是由于借鉴了成熟的方法。但其是否真正适合团队的实际情况，却没有评估。当我们发现这种方法的弊端时，就难以再按照原路返回了。所以对于一些看似很好的“别人的经验”，要慎之又慎。

14.1.4 平衡是一个难题：萨伊定律与凯恩斯法则

明确了工作的流程和工作的方式，接下来就是期待工作的产出，毕竟产出代表了工作的价值所在。作为研究人员，一侧受研究部门的管理，另一侧也要受到业务单元的约束。这便天然地产生了一对矛盾，研究是一种关注长期的工作投入，有的时候很难衡量产出，甚至由于试错而没有产出；而事业部则更倾向于关注产品化，要的就是马上生产获得收益。夹在两者中间一定很难受，这个时候就需要平衡。

法国人让·巴蒂斯特·萨伊（Say JeanBaptiste）是18世纪后半叶最为知名的

经济学家。他创造了属于他自己的理论“萨伊定律”，这个理论认为，只要供给充分，就可以产生足够的需求，于是卖可以产生买，经济就会好转。通过他的分析，经济显然并不需要干预。

而 100 多年后的英国经济学家约翰·梅纳德·凯恩斯（John Maynard Keynes）则提出了截然不同的观点。他认为，为了使经济回到正轨，需要刺激买的需求，这样才能够拉动生产者生产。透过他的分析，经济需要被干预。

本质上来说，两者围绕的正是经济中究竟应该看长期还是看短期的争执。短期看，凯恩斯占优；而长期来看，萨伊定律有理。看来我们在讨论工作的时候，也需要讲究长期和短期的平衡。平衡本身就是一个讲究场合、场景甚至艺术的话题，知易行难。

14.2 人员之间的协同

事情做好了，却不一定代表做对了，往往用于衡量事情的是结果的质与量，而我们却忽略了对过程中人之得与失的关注。无论是团队内的人员协作，还是团队间的人员协作，我们都要用“利他”的胸襟来包容。

14.2.1 你闪开，让我来：绝对优势与相对优势

当我们只是团队的一员时，我们的职责就是把手头上的工作出色地完成，并尽可能帮助领导承担更多的事务。一旦我们的工作内容从自己做事变成带他人做事，职责就绝不是完成手头工作那么简单了。对他人的培养需要额外关注。

然而，对于很多新晋的 Leader 来说，容易走两种极端。一种是完全甩手，另一种是不肯放手。前一种情况源于责任心的缺失，后一种情况则出于个人性格特点。当团队成员没有完成交代的工作时，Leader 心中升起一股怨念，心想“这么简单的事情都不会做”，于是就出现了这样的一幕——你闪开，让我来。

我们在技术转型管理的过程当中，往往有一种惟我独尊的心态，这种心态本身没有问题，但是我们只关注了自己的绝对优势。作为团队 Leader，招一个新人进来做事，就要做好最坏情况的应急方案。如果 Leader 花时间帮助新人完成工作，

意味着 Leader 失去了一份本可以完成自己工作的时间，团队成员也因此失去一次自我成长的机会。结果是自己越来越忙，团队越来越闲。

绝对优势最早是由 18 世纪英国古典经济学家，《国富论》的作者，亚当·斯密（Adam Smith）提出的，用来阐述分工的必要性，也为自由贸易奠定了理论基础。

相对优势则是同为英国经济学家的大卫·李嘉图（David Ricardo）在其著作《政治经济学与赋税原理》中提出的。他认为“两利相权取其重”，某个国家和地区在多种产品的生产上可能相对其他国家与地区都是有优势的，因此在选择生产时要使用机会成本来判定生产哪种产品在哪方面有相对优势。

以上面 Leader 和新人的例子说明，Leader 完成一个计划书的编写可以为团队带来 100 万元的收益，而编写程序可以带来 1 万元的收益。对于新人来说，写计划书没有经验，也许只能创造 1 万元的收益，而编写程序则可以带来 1000 元的收益。从商业计划书和程序编写两个任务上来看，Leader 比新人都有绝对优势。但是对于 Leader 来说，放弃写计划书而选择写程序，等同于放弃了 100 万元的机会成本。而对于新人来说，写程序所放弃的机会成本仅为 1 万元，其机会成本更低，因而在写程序这件事情上，新人比 Leader 更有相对优势。

14.2.2 无条件开放：零和博弈与合作共赢

团队工作不仅需要有自己的关注，还要关注与其他团队的配合。在研究与业务并不紧密耦合的公司内部，研究部门关注的是模型与算法引擎的研发，事业部则侧重于业务与产品的收益。两个部门是不同的团队，从研究部门角度看，双方团队在配合上有两种模式。

一种是“大包大揽”的模式，另一种是“无偿交割”的模式。在前者的模式中，研究团队既做研究，也做工程。这样做的好处是能够快速落地，对于与之合作的事业部来说，也可以快速拿到产品转化成销售。但久而久之，事业部的研发人员并不能够胜任这样的工作，使得能力停滞不前。这种模式很容易使研究团队停留在“舒适区”，他们会担心这样的工作交给事业部之后，事业部会培养出能够独当一面的研发队伍，从而把核心技术对接走，造成“抢饭碗”的局面。

与之相反的是另一种称为“无偿交割”的模式。在这种模式下，研究部门将自己的成果对事业部进行技术开放，甚至毫无保留地将模型与算法交接给他们。这个过程看似是在缩小双方团队的差距，但实际上却是在帮助公司逐步做大。从大局观上说，我们并不是为了自己而工作，而是奔着公司的大目标前进与努力。

我个人倾向于选择后面的工作模式，甚至为了使事业部团队接手更方便，还要对其进行培训，帮助其尽快将队伍建立起来。工作在很多时候并不是自己或所在团队的独角戏，其更像是一场足球，大多数情况我们都在补别人的位。与其把别人罚下场，不如帮助队友提高水平，大家踢一场高质量的球赛。

经济学中的“零和博弈”与“合作共赢”讲的就是这样两种不同的工作模式。“零和博弈”的概念源于博弈论，博弈论萌芽于 20 世纪 20 年代，1944 年由冯·诺依曼（John von Neumann）与美籍经济学家奥斯卡·摩根斯特恩（Oskar Morgenstern）联合提出。“零和博弈”指的是，在对弈情况下，一方获益必然导致另一方损失，对弈各方的整体收益并没有增加。与之相对应的是“合作共赢”的模式，即双方都能够从博弈过程中有所收获，达成“双赢”。

14.2.3 教会团队成员什么是沉没成本

记得以前看过央视的一个栏目《挑战主持人》，其中的一期节目让我印象深刻。当时一个参与节目的选手在上台发言时不小心被脚下的台阶绊了一跤，摔倒了。一个本是用来展示良好形象的舞台变成了一个让自己窘迫的情境。选手在爬起来之后说了一句让我难忘终身的话，“从哪里跌倒，就从哪里爬起来”，赢得了现场阵阵掌声。

工作中也会出现类似的糟糕情况，在这种情况下，如何处理则决定了事态发展的方向。还记得我们团队的一个实习生在入职的第一天弄坏了公司的一台服务器，我想当时他的心中应该是忐忑的。如果这个时候去计算损失并且加以指责，不仅挽回不了损失还有可能对其造成心理阴影，不利于未来的工作开展。这个时候需要做的事情有两件：一是让错误发生的当事人以正式的邮件方式向受到其影响的人说明情况（注意，不是道歉），影响程度以及解决方案；二是通过团队内部

复盘找到根本问题，补足流程。前者是为了让他记住这件事，后者是为了预防此事再次发生。

那些已经失去的事物或者犯下的错误就好比一只摔在地上的甜筒或是一张丢失的电影票，已经是不可挽回的了。在经济学中有一个特定的名词用来形容那些已经发生且不可收回的支出，称为“沉没成本”。人们在做决策的时候，不仅要看未来的收益，也要看过去的投入。例如很多人大学毕业后发现自己的专业并非热门专业，也非上升势头，明明有机会选择其他的职业重新开始，获得光明的未来，但是由于舍不得大学四年在本专业上所花的时间，因此错失发展机会。

第 15 章

软实力：靠心理学打造

- 15.1 向内求：耐心、谦逊、热心..... 294
 - 15.1.1 让自己“延迟满足” 294
 - 15.1.2 对表扬免疫 295
 - 15.1.3 不怕丢脸地分享 297
- 15.2 对外看：大局、妥协、有趣..... 297
 - 15.2.1 看问题需要“上帝视角” 298
 - 15.2.2 率真对内，圆滑对外 298
 - 15.2.3 一切从简，有趣有梦 299

15.1 向内求：耐心、谦逊、热心

什么才是产品经理最重要的软实力呢？在我的认知中，对事情耐心、对他人谦逊，以及对社会热心是最为重要的三个品质。

15.1.1 让自己“延迟满足”

吃过美式中餐的人应该对其中的幸运饼干印象深刻。这是一种折叠起来的小饼干，饼干中藏有一张充满了励志箴言的纸条。图 15-1 是我在 UCLA 学生餐厅 Panda Express 用餐时吃到的一张字条，上面写着：“You will be rewarded for your patience and understanding”（你将会因为耐心和理解受到奖赏）。“耐心”与“理解”作为衡量自控力的重要品质之一，引起了我的注意。



图 15-1 UCLA 校园餐厅供应的幸运饼干

在心理学中，有一个著名的“儿童糖果实验”，或称为“棉花糖实验”。该实验是由斯坦福大学人格心理学家沃尔特·米歇尔（Walter Mischel）于 20 世纪 60 年代进行的。这个实验最早是在斯坦福大学的幼儿园中进行的。研究人员找来数十名儿童，让他们独自待在一个仅有桌椅的房间中。桌上摆放着棉花糖、糖果等儿童喜爱的食物。研究人员给他们两个选择，可以马上吃掉零食，或者

等研究人员回来再吃。如果等到研究人员回来再吃，可以获得多一倍的零食作为奖励。

对于儿童受试者而言，实验过程十分煎熬。许多孩子没到 3 分钟就把零食吃光了，只有三分之一的孩子等待了 15 分钟，等来了研究人员以及多一倍的奖励。后来实验进一步扩大，米歇尔持续跟踪了这些受试者多年，结果发现，那些在孩童时期能够等待研究人员归来从而获得多一倍奖励的受试者在学生时期以及走上社会后，往往能够表现得更为出色。这种愿意为了未来或者长期价值而甘愿放弃当下诱惑和即时满足的抉择取向，被称为“延迟满足”。即时满足还是延迟满足，对于受试者来说是二选一的，这也是个人自控力的体现。

无论是 UCLA 幸运饼干的警句，还是心理学中的“儿童糖果实验”，它们仿佛都讲述了“延迟满足”的道理。“延迟满足”并非是要人们压抑欲望，也不是要人们忍受痛苦，而是锻炼一种克服当前困难从而谋求长远利益的韧劲与心态。

冲动时的行为尽管可以图一时之快，却有可能引一世之殇。数据产品经理会时常感到自己如履薄冰、战战兢兢。我们既要承受无授权领导的质疑，也要背负团队 KPI 的压力，往往在产品的短期交付和长期规划中进行平衡。在这个过程中，能否让自己“延迟满足”，不断培养自己的战略耐性，显得尤为重要。

不同的职业与个人亦有着不同的职业分布曲线。如果把一生的成就看成曲线与坐标轴围成的面积，有的人的曲线是正态分布，即青年时上升，中年时下降；有的人的曲线是长尾的指数增长，即在中年才开始爆发。两者对比看来，累计面积并无区别。但若是有着后者职业曲线的人在青年时就没有了耐心，不能够“延迟满足”，便是没等到磨砺好就宝剑出鞘，最后只能“画虎不成反类犬”。

15.1.2 对表扬免疫

日常生活中我们常能听到别人在应对表扬时答道“哪里，哪里”。其实受表扬人内心是尴尬的。心理学研究表明，人们受到表扬时的心态往往有两种。一种是自喜，这种情况大多发生在别人对自己的评价与自己的评价吻合时。另一种是自卑，受表扬者要么觉得自己配不上这样的评价，要么感觉表扬是他人对自己未来的期待从而产生压力，或是感受到了来自表扬的权利结构性压迫致使产生逆反心

理。既然表扬有这么多副产品，那么作为接受者究竟该如何应对呢？

二战时期，为了预防出现纳粹德国侵占英国的情况，英国政府印制了一系列海报用于鼓舞民众的信心和军队的士气。其中最为出名的一张，就是众所周知的“Keep Calm and Carry On”（保持冷静，继续前行）。



图 15-2 英国政府印制的海报

图片来源：<https://baike.baidu.com/item/Keep%20Calm%20and%20Carry%20On/7206538>

正如我们在前文提到逻辑斯蒂回归时介绍物种群体增长的 S 型线一样，我们个体的职业发展也是同样的情况。开始，斗志昂扬；中途，些许疲惫，增长乏力；最后，发展停滞，职业倦怠。

包括数据产品经理在内的大多数行业人物角色设定似乎就是这样的一条 S 型曲线。因此，为了突破限制与瓶颈，需要突破渐进性、连续型增长，以另一条技能和发展的 S 型曲线取而代之。两条 S 型曲线连接之处的鸿沟便需要我们靠清醒的头脑与持续的积累去跨越。

归根结底，能否克服个人、企业、社会的不连续性，在于我们是否能够找到下一条增长曲线，不停地自我革命和自我竞争。而在此过程中保持清醒的头脑，放下已经取得的成就，对获得的表扬免疫，这些至关重要。

15.1.3 不怕丢脸地分享

在我上初中的时候，疯狂英语的创办人李阳非常出名。我并没有学过他 CD 中的口语课程，但是视频中的一个画面却让我印象深刻。李阳为了帮助学员走出“哑巴英语”的窘境，克服与人交谈时因为发音不准被人嘲笑的面子问题，带领一众学员在大街上跑步，边跑边喊出了他的那句精神引领性质的口号：“Don't be afraid of losing face”（别害怕丢脸）。

2017 年的毕业季是最为精彩的时节，许多知名大学邀请政商学界名人来到毕业典礼的现场为毕业生们送去步入社会前的建议与良言。星巴克公司的 CEO 舒尔茨被邀请在亚利桑那州里大学（Arizona State University）发表演讲。在演讲中，他提出了三个需要大家常念的问题，其中之一就是“如何不失尊重地与他人分享自己的成功”。对于很多人来说，成功远远谈不上，但分享却可以轻松做到。

工作中的很多同事不愿意分享，究其原因主要有二。一是小气，二是怕丑。小气是害怕自己的优势被他人学去，怕丑则是觉得自己知道的东西没有什么好分享的，在别人面前分享他人知道的事情是“哗众取宠”。

无论是上述两种中的哪一种，都是害怕与人建立连接。我们曾在介绍搜索引擎的网页排序算法时从信息学的角度出发，给出了网页能够流行的几个特质。这些特质与社会中信用背书的三种方式相似，即“你行，有人说你行，说你行的人还得行”。分享就是不停地将自己接入系统，只有当连接足够多，且连接的质量足够高时，作为系统和网络中的一员才能获得高质量的评分。

15.2 对外看：大局、妥协、有趣

在柏拉图的理想国中，尽管大家都遵从“金属人”的童话信仰，但是理想国的统治者却既需要向内修炼体育、音乐与哲学，也需要走下神坛，走入这毫无偏袒的世界。因此我们除了向内求，对外看也很有必要。拥有大局观，学会“妥协”以及一切从简，是和周遭打交道的最好方式。

15.2.1 看问题需要“上帝视角”

“唐宋八大家”之一的苏轼曾在由黄州赴汝州上任途中，游览庐山，留下了千古名句“不识庐山真面目，只缘身在此山中”。后人对这句诗的解读往往偏重于“当局者迷，旁观者清”，但谁又知背后的深意。苏轼由于“乌台诗案”被贬黄州，后来奉诏调任汝州，本是东山再起，是喜。谁料上任途中却遭遇了丧子之痛，是悲。情感的起与落，使得苏轼在面对庐山时发出这样的感叹，不免有“造化弄人，身不由己，看不清，道不明”之感。

工作与生活过程中，难免有苏轼这种置身局中，不得脱身的尴尬。正如苏轼渴望跳脱出来看整个朝廷政局一样，我们也希望能够站在一个第三者的视角上重新审视自己所从事工作的意义、价值，思考是否有更好的方法，更清晰的方向。这样抽身事外，看清事物本质的视角就称为“上帝视角”。

猎豹 CEO 傅盛曾从《三体》中得到灵感，于是提出商业竞争中的要领是“升维思考，降维打击”。他也曾引用苹果前 CEO 约翰·斯卡利（John Sculley）的观点，即经营公司既需要 Zoom out（抽象），也需要 Zoom in（聚焦）。前者指的是宏观层面的运筹帷幄，后者指的是细节层面的“锱铢必较”。无论是升维思考还是 Zoom Out，其实质都是建立宏观的思维能力，亦是一种“上帝视角”。

对于数据产品经理来说，每天除了和原型设计、需求评审、数据预处理等工作细节打交道之外，也需要参与产品未来方向的规划与设计，参与部门乃至公司的业务架构调整。前者处理细节，容易让人“身陷囹圄”，焦头烂额；后者处理大局，则需要置身事外，审时度势。在这种情况下，谁能够及早抽身出来看问题，具备大局思维和战略思考能力，谁就能够有更长远的发展。

15.2.2 率真对内，圆滑对外

以前我总以为率真与圆滑是一对反义词，后来才发现两者其实是一个硬币的两面。圆滑对外，给人尊敬，让人舒适，打磨的是我们与外界的交互能力。率真对内，让己从容，坚持初心，锤炼的是自己言出必行的作风。

数据产品经理是一个与人打交道的岗位，尽管我们忠于自己的产品，但也要

懂得在客户价值与公司利益面前“曲线救国”，适当妥协。尽管产品经理们把乔布斯当作偶像，把他的“Follow Your Heart”奉为圭臬，但很显然一些资历较浅的产品经理所追随的不过是个人喜恶的初心，而不是目标达成的初心。

我们常说要让领导做选择题，而不是填空题，但许多人觉得给出的选择题只有一个选项，这个选项就是自己喜欢的选项，是自己当下处理问题最为省事的选项。有一种很简单的技巧可以化解这样的矛盾。在给出一个自己偏好的选项后，再找出一个与之对立的可行性方案。讨论问题并非说服周边人接受自己的观点，而是要说服大家接受某一个可行观点，从而团结力量，劲往一处使。

妥协，不是示弱，而是帮助别人，成就别人。一位在澳洲某大学任教的老师曾告诉我：“帮别人就是在帮老天，老天也会帮助你”，这种“成己为人，成人达己”的心态，需要包括我在内的每一位产品经理不断修炼。

15.2.3 一切从简，有趣有梦

设计学上有一个经典问题称为“前后台简单选择问题”。后台指的是与制造产品工程师们打交道的产品部分，而前台则是指与使用者打交道的产品部分。

后台简单而前台复杂的系统往往在设计与制造的时候会按照简单、标准和模块化的方式生产，而把功能的组合、操作与使用的任务交给用户。乐高玩具就是这种类型，每个模块都很简单，但组合方式复杂，很显然这种设计方式适合发烧友群体。

与之相反的，后台复杂而前台简单的系统则是把杂乱无章的组合隐藏在背后，前台与用户的交互只有用户想要的若干个功能。各类遥控器以及一体台式机便属于这种类型，你并不需要关心启动一台机器需要在内部执行多少项检查，先后执行什么模块，只需一个按键便可以搞定一切，这种设计方式适合实用类型的产品。对于大多数事物，我们都希望它们能够尽可能简单。

无论是 MUJI 还是 SONY，日系产品总能给你简约大方，清新脱俗的感觉。将简单发挥到极值的“断舍离”由日本整理术大师山下英子创立，从物质与心灵的角度给出了极简生活的要领。与之相反的则是具有收集癖好的“松鼠效应”。我们收藏微信文章，办各类会员卡，攒各种小物件，就像一只辛勤的松鼠，不停地

向自己的洞中搬运食物，而真正食用的却很少。

我们本不擅长复杂，但见的世面多了，便也学会了从众，认为只有复杂才能应对复杂，然而殊不知简单才是最容易实践的，是维持一致性最好的方式。它既让别人在与我们协作时预测我们的行为从而心安，也能化解我们个人生活中的复杂。